K-3 Reading Assessments

Scott Paris
University of Michigan
With the collaboration of Rob Carpenter, Alison Paris, David Pearson, and many colleagues at CIERA

Center for the Improvement of Early Reading Achievement

www.ciera.org
Reading Realities in 2003 in K-3

Classroom Practices

• Focus on basic skills
• Less emphasis on comprehension & variety of genre
• Repeated assessments & early accountability

National Policies

• Endorsed assessments
• Preferred materials
• Scientifically based research guides policies
Today’s Problems

• Teachers are unhappy with skills emphasis
• Students endure drills & testing
• Parents want evidence of high and early achievement
• Policies driven by accountability
• High ground of research is held by NICHD researchers & basic skills
• Reading First legislation distributes $5B and dictates classroom practices
Consequences for Early Reading Assessment

• New state-designed tests of early reading
• The 5 Big Ideas in reading prescribe what is assessed
• The use of IRIs for summative and formative assessment
• Measures of AYP required in grades 1-5
• Must be based on scientific research but research is contested
State-Designed Early Reading Assessments

• Michigan Literacy Progress Profile (MLPP)
• Texas Primary Reading Inventory (TPRI)
• Illinois Snapshots of Early Literacy (ISEL)
Michigan Literacy Progress Profile (MLPP)

• Designed for K-3; is being expanded to K-5
• 5 Milestone tasks (Taking Stock)
• 6 Enabling tasks (Digging Deeper)
MLPP Milestone Tasks

- Oral language
- Oral reading fluency
- Comprehension
- Writing
- Attitudes & Self-Perceptions
MLPP Enabling Tasks

- Letter-sound identification
- Phonemic awareness
  - Rhyme choice & rhyme supply
  - Blending
  - Segmentation
- Concepts about Print
- Hearing and Recording Sounds
- Known Words
- Sight Words/Decodable Words
Reliability & Validity of MLPP

- UM - MSU-Ingham ISD collaboration with MDE the past 4 years
- Required many years to conduct the research
- Data support MLPP assessments
- Surprising new ideas about reading skills
Test-retest correlations for the MLPP enabling skills

- Letter Identification: .96
- Letter-Sound identification: .86
- Phonemic Awareness: .93
- Hearing & Recording Sounds: .93
- Known Words: .73
- Concepts of Print: .56
Conclude

• Enabling skills can be assessed reliably, especially phonics and phonemic awareness

• Concepts of Print and Known Words may be less stable due to learning during testing or examiner effects
**Test-retest correlations for the MLPP oral reading skills**

- Based on QRI, DRA, and BRI
- Collected from readers in grades 1,2,3
- Averaged over highest and lowest passages read by each child, i.e., independent and instructional levels
Test-retest correlations for the MLPP oral reading skills

Sight Word Identification  .70
Oral reading rate  .82
Oral reading accuracy  .45
Oral reading retelling  .53
Oral reading comprehension  .58
Conclude

• Oral reading assessments are less reliable than enabling skills, perhaps because of learning/familiarity during testing or examiner effects

• Reading rate is most reliable, followed by word recognition, suggesting that they may be more stable skills
Concurrent Validity of MLPP Enabling Tasks

Grades K & 1; correlated with similar tasks in TPRI

Letter identification .92
Letter-Sound identification .54
Phonemic Awareness .77
Correlations of Oral Reading Between MLPP and TPRI

Averaged over highest and lowest passages

Sight word identification  .79
Reading rate  .77
Accuracy  .46
Comprehension*  .27

*Possible ceiling effect in 5 item TPRI comp task
Concurrent Validity of MLPP Enabling Tasks with GMRT

<table>
<thead>
<tr>
<th>Task</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter identification</td>
<td>.82</td>
</tr>
<tr>
<td>Letter-Sound identification</td>
<td>.55</td>
</tr>
<tr>
<td>Concepts of Print</td>
<td>.45</td>
</tr>
<tr>
<td>Phonemic Awareness</td>
<td>.19</td>
</tr>
</tbody>
</table>
Concurrent Validity of MLPP Oral Reading Tasks with GMRT
Averaged over highest and lowest passages

Sight word identification .91
Propositions recalled .70
Comprehension .77
Predictive Validity of MLPP Enabling Skills with GMRT

<table>
<thead>
<tr>
<th>Enabling Skill</th>
<th>6 mos</th>
<th>1 yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter identification</td>
<td>.46</td>
<td>.40</td>
</tr>
<tr>
<td>Letter-Sound identification</td>
<td>.73</td>
<td>.68</td>
</tr>
<tr>
<td>Total Phonemic Awareness</td>
<td>.60</td>
<td>.64</td>
</tr>
<tr>
<td>Sight Word identification</td>
<td>.89</td>
<td>.84</td>
</tr>
<tr>
<td>Hearing &amp; Recording Sounds</td>
<td>.68</td>
<td>.73</td>
</tr>
<tr>
<td>Concepts of Print</td>
<td>.43</td>
<td>.34</td>
</tr>
</tbody>
</table>
Predictive Validity of MLPP Oral Reading Skills with GMRT

Oral Reading Skill          6 mos  1 yr

Accuracy                    .67   .74
Propositions Recalled       .65   .61
Comprehension               .58   .60
**Predictive Validity of MLPP Skills with MEAP**

Averaged over highest and lowest passages

<table>
<thead>
<tr>
<th></th>
<th>1 yr</th>
<th>2 yrs</th>
<th>3 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sight words (District C)</td>
<td>.65</td>
<td>.52</td>
<td>.47</td>
</tr>
<tr>
<td>Accuracy (District C)</td>
<td>.62</td>
<td>.36</td>
<td></td>
</tr>
<tr>
<td>Accuracy (IRT scores)</td>
<td>ns</td>
<td>ns</td>
<td>.30</td>
</tr>
<tr>
<td>Propositions recalled</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.40</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td>Comprehension (IRT)</td>
<td>.37</td>
<td>ns</td>
<td>.21</td>
</tr>
</tbody>
</table>
Conclude

• Enabling skills have higher correlations with similar skills on other tests so greater concurrent and predictive validity

• Oral reading accuracy, comprehension, and retelling may be more variable across texts so lower correlations

• Both enabling and milestone MLPP skills demonstrate reasonable validity
Surprising Insights

1. Most enabling skills are mastered in K-1 so their distributions change with age.
2. Skewed distributions have dramatic effects on correlations used to establish validity.
3. Exp. Letter recognition and COP show good correlations at K but lower correlations at 1st grade after mastery.
Surprising Insights

4. Constrained skills develop early and quickly
   - Letter identification
   - Letter-Sound identification
   - Concepts of Print
   - Hearing and Recording Sounds
   - Phonemic Awareness

All yield transitory correlational patterns
Surprising Insights

5. Many tasks assess automatic decoding
   Sight word recognition
   Oral reading rate
   Oral reading accuracy

These may reflect more stable individual differences than comprehension and recall
Surprising Insights

6. Comprehension and vocabulary skills develop continuously and are less constrained in distributions

7. But neither is assessed as often or in as much detail as decoding skills.
Texas Primary Reading Inventory (TPRI)

Designed for K-2 assessment of:

- Book and Print Awareness
- Phonemic Awareness
- Oral Reading Accuracy & Fluency
- Listening & Reading Comprehension

- 9 short grade level passages w/5 Qs each
Illinois Snapshots of Early Literacy (ISEL)

1. Alphabet recognition
2. Story listening
3. Phonemic awareness
4. One-to-one word naming and matching
5. Letter sounds
6. Developmental spelling
7. Word recognition
8. Graded passage reading
ISEL Graded Passages

• 3 for first grade, one for second grade
• Passages increase from 40-100 words
• Teachers record miscues, fluency as intonation, and answers to 2 questions
MLPP, TPRI, & ISEL Are

- Child-focused & teacher controlled
- Tied to professional development
- Designed for individual, diagnostic, early assessment of multiple reading skills
- Not easily used for summative reporting
- Strong on decoding and enabling skills
- Weak on comprehension assessment
- Susceptible to misinterpretation of constrained skills & prescriptive errors
Prescriptive Error

- Child scores poorly on letter recognition
- Therefore, child needs more instruction on letter recognition.
- This error leads to repeated assessment and instruction on enabling skills when these are only indicators of slower mastery that might be due to many other factors
- Thermometer analogy
Consider Alphabet Knowledge

- **Claim**: Lonigan, Burgess, and Anthony (2000) said, "... knowledge of the alphabet (i.e., knowing the names of letters and the sounds they represent) at entry into school is one of the strongest single predictors of short- and long-term success in learning to read ... " (p.597).

- **Implication**: Teach and assess the skill early to insure reading success.

- **Question**: Is this a valid inference for practice?
Many Early Reading Assessments Focus on Constrained Skills

- Comprehensive Test of Phonological Processing - CTOPP
- CBM Oral Reading Fluency
- Woodcock Reading Mastery Test
- DIBELS
The DIBELS Assessments

- Dynamic Indicators of Basic Early Literacy Skills
- Created by Univ of Oregon researchers
- Can be downloaded for free; website includes video clips and background info
- http://dibels.uoregon.edu/dibels
- Designed as series of 1 minute assessments
- Measures 3 of 5 Big Ideas, not vocabulary or comprehension
DIBELS Measures

• Phonological Awareness
  – Initial Sounds Fluency
  – Phonemic Segmentation Fluency

• Alphabetic Principle
  – Nonsense Word Fluency

• Fluency
  – Oral Reading Fluency (wcpm)
My Concerns With the DIBELS

• One-minute samples are narrow and shallow instances of reading skills
• Emphasis on automatic word recognition is good but emphasis on rapid reading is bad
• Too much testing on enabling skills and not enough assessment of comprehension
• Prescriptive model is based on transitory correlations of constrained skills
Let’s reconsider the 5 Big Ideas

1. Alphabetic knowledge - highly constrained
2. Phonemic awareness - constrained
3. Oral reading fluency - constrained

All necessary but not sufficient for proficient reading
4. Vocabulary - unconstrained
5. Comprehension - unconstrained
Implications of Constrained Skills

- Skills exhibit developmental asynchronies
- Correlations vary according to skew in data and sample characteristics
- Unstable correlations reflect conceptual & measurement constraints
- Predictive validity claims are transitory
- Correlations are confounded with many concurrent developments
- Proxy correlations are misinterpreted as causes of reading problems
A Deceptive Proxy Variable

- Reading rate = wcpm
- Rate increases within Ss with age and expertise
- Faster readers have higher achievement scores
- Fast reading is a proxy for automatic decoding and other skills
- No data to show that increasing reading rate enhances comprehension
Reading Rate Is a High Stakes Measure

CBM (Fuchs & Fuchs)
DIBELS (Good, Simmons, & Kame’enui)

Why?

• Quick, easy, efficient, repeated testing
• Used with any text
• Yields quantitative data
• Correlates highly with achievement data
Reading Rate

- Increases 10-15 wcpm per year in K-5
- Restricted range & variance by age
- Proxy for automatic decoding
- Assumes rate is independent of text features, vocabulary, prior knowledge, & motivation
- Confounded correlations between-tasks and between Ss
Problems With a Focus on Reading Rate

- One-minute samples are artificial tasks
- Rate is a narrow measure of reading
- Speed is the wrong purpose of reading
- Sends wrong message to children & parents
- Confounded interpretations of rate correlations and prescriptive errors
- Bogus claims about predictive validity
- Detrimental assessment/instruction policies
Oral Reading Fluency

• Assessed with running records or miscue analyses
• Frequent in grades 1-2
• Includes all 3 components below but often assessed only with rate
  – Accuracy
  – Reading rate
  – Intonation (prosody)
Informal Reading Inventories (IRIs)

- Examples: QRI, BRI, DRA
- IRIs can include:
  - Oral reading accuracy measures
    - Miscues or running records
  - Fluency ratings
  - Reading rate
  - Retelling
    - Rubrics or propositions or main ideas
  - Comprehension questions
    - Implicit and explicit questions
  - Graded Word lists
Oral Reading Measures

Benefits
- Aligned with daily instruction
- Multiple measures of fluency & comprehension collected simultaneously
- Diagnostic immediately

Liabilities
- Requires expertise to administer & interpret
- Requires multiple passages to assess reading level
- Accuracy (i.e., running records & miscues) are insufficient by themselves
- Teachers may “teach” the commercial materials
Need to Collect Information

- From multiple passages & levels
- From multiple genres
- From multiple word lists
- From silent reading for children reading at grade 4 and above because
  - Comprehension and accuracy unrelated for many children
  - Silent reading allows look backs & strategies
  - Silent reading avoids social anxiety
Concerns About Fluency Assessments

- Focus on rate alone is wrong, i.e., CBM
- Prosody is difficult to assess reliably, shows little change on 4 point rubrics, and carries less variance than accuracy and rate
- Accuracy is skewed above 90-95% and the relation between poor accuracy and poor comprehension is stronger than high accuracy and good comprehension
Caution: Fluency Does Not Mean Good Comprehension

- **Word callers** - High accuracy, low comprehension
- **Gap fillers** - Low accuracy, high comprehension
- More of both kinds of readers after grade 3 so silent reading and comprehension assessments are needed
Figure 1. Pretest Correlations Between Oral Reading Factor and Comprehension Factor

![Graph showing pretest correlations between oral reading and comprehension factors across different passage levels. The graph indicates a generally decreasing trend in correlations as passage level increases.]
Why do correlations between oral reading accuracy and comprehension decrease with increasing passage difficulty?

Why are the correlations so low in general?
The Fluency-Comprehension Disjunction Hypothesis

• Accurate oral reading and comprehension may be positively correlated best ($r = .4-.6$) when:
  – Text is easy to decode
  – Text is brief and familiar
  – Questions are factual or easy inferences
  – Assessed among young/beginning readers
The Fluency-Comprehension Disjunction Hypothesis

- Oral reading accuracy may become less related to comprehension among:
  - Older readers in grade 3 and above
  - More skilled and faster decoders

Because automaticity of decoding provides necessary but not sufficient foundation for understanding
Implications

- Measures of oral reading accuracy may be limited beyond grade 3 to assessments of word identification.
- Need to compare oral and silent reading comprehension as functions of text difficulty, reading skill, and age.
- Need to recognize that slow, inaccurate readers can be good comprehenders and fluent readers may not be good comprehenders.
IRIs Are Diagnostic When Teachers

- Interpret patterns of oral reading miscues & self-corrections
- Identify difficulties answering specific questions or retelling information
- Use results for conferences with children (e.g., retrospective miscue analysis)
- Align reading materials and instruction with children’s needs
IRIs May Present Problems

- When focus is only on rate or accuracy
- When not connected to instruction
- When not given by teachers
- When data are used as summative or comparative assessments
Reporting Achievement and AYP With IRIs

Problems:

• Difficult data to collect reliably
• Different passages and levels: How can they be compared?
• Skewed data with narrow range does not allow much room for improvement in accuracy or rubrics for retelling
• What is significant growth?
Six Potential Solutions
(Paris, Reading Teacher, 2002)

1. Same texts for repeated testing
2. Selected, scaled texts
3. Categorical, criterion-referenced reports
4. Weighted calculations of text difficulty
5. Item response theory (IRT)
6. Hierarchical linear models (HLM)
Identical Texts

- Avoids level & passage confounds
- Can use any texts over multiple tests
- Easy statistical analyses
- Practice effects
- Limited text sample
- Limited range to demonstrate growth
Selected, Scaled Texts

- Levels & difficulty specified
- Range of genre & difficulty specified
- Texts linked across test times

- Scales may be debatable
- Scales may be different for different skills
- Cost of design of comparable texts
**Criterion-Referenced Reports**

- Usually reported as Independent, Instructional, and Frustration levels
- Allows comparable performance categories across texts
- Matches teachers’ use

- Establishing criteria is difficult for multiple skills & texts
- Less detailed diagnostic information
- Progress defined by two scales of criteria & text difficulty
Weighted Difficulty Scores

• Multiply performance score by difficulty scale such as Lexile, DRP, or readability units
• Yields single scale across text levels
• Easy analyses

• Debatable difficulty scales may not apply equally across skills
• Uncertain how to weight scores
• Are weights applied the same way across texts, genres, & ages?
IRT scaled scores

- Creates single scaled scores across texts & levels
- Can be used with any & many texts
- Clear interpretation of results

- Requires large sample, good fit of data, and sophisticated statistical analyses
- Not detailed & diagnostic
- Turns IRIs into high-stakes tests
Possible Solutions

All can be aggregated by classroom and school and reported as Gain/No Gain or Text Level/Standard met.
Conclusions About IRIs

- IRIs are reliable and valid assessments
- IRIs can serve both formative and summative functions
- IRIs and MLPP assessments can provide a solid foundation for PD
- Early reading assessment should reinforce instruction & curricula
How Can We Assess Comprehension In Children Who Cannot Decode Text?

Important to Balance Assessments of Decoding Skills with Comprehension

- Listening comprehension
- Story telling & retelling
- Picture Book task to assess narrative comprehension

- Authentic measure of emergent readers’ narrative comprehension skills
- Does not require decoding skills
- Links assessment and instruction
- Picture Walk
- Retelling
- Prompted Comprehension
ROBOT-BOT-BOT

Fernando Krahn
NC Task Part 1: Picture Walk

- Observe five types of behaviors
  1. Book handling skills
  2. Engagement
  3. Picture comments
  4. Storytelling comments
  5. Comprehension strategies

- Scoring
  - 0-1-2 point rubric
  - 0-10 point Picture Walk scale
Part 2: Retelling

- Story elements scored in retellings:
  1. Setting
  2. Characters
  3. Goal/initiating event
  4. Problem/episodes
  5. Solution
  6. Resolution/ending

- Scoring
  - 0-6 point Retelling scale
Part 3: Prompted Comprehension

Explicit Information
- Setting
- Characters
- Initiating event
- Problem
- Outcome resolution

Implicit Information
- Feelings
- Causal inference
- Dialogue
- Prediction
- Theme
NC Task Research Shows

• Developmental improvement with age
• Readers score better than nonreaders
• Easy and reliable to administer
• Significant concurrent validity
• Significant predictive validity
Big Picture Issues About Assessment: Students

Surface Issues
- Classroom Tests
- SAT, ITBS
- Report Cards
- Conferences
- *Am I smart?*

Deep Issues
- Reflect & Monitor
- Set Goals
- Self-assess
- Self-efficacy
- *Self-regulated, motivated learners*
Big Picture Issues About Assessment: Teachers

**Surface Issues**
- Curriculum assessments
- High-stakes tests
- Test preparation
- Grading
- Report cards
- Conferences
- *Too many tests, not enough time*

**Deep Issues**
- Daily diagnosis of students
- Aligning instruction with assessment
- Authentic evidence
- External pressure for accountability
- *Using assessment to support instruction*
Big Picture Issues About Assessment: Parents

Surface Issues
• Report cards
• Conferences
• Test scores
• How does my child compare to others?

Deep Issues
• Communication
• Homework & support
• Learning goals not grades & scores
• Personal, developmental analysis of progress
Big Picture Issues About Assessment: Administrators

Surface Issues

• Compliance & Collection of data
• Standards achieved
• Media reports
• Costs
• Public accountability & status

Deep Issues

• Alignment of assessment, curriculum, and instruction
• Staff development
• Reconciling different agendas of policymakers, teachers, parents
• Achievement, staff development, values, jobs