The Problem

Research tells us that some academic interventions and intervention approaches work better than others (on average)...

But how do we know if an intervention is working for a particular student?

[and what do we do when it isn’t working?]
Purpose

• Introduce Data-Based Individualization (DBI) for service delivery

• Introduce Curriculum-Based Measurement (CBM) as a data source

• Discuss challenges & solutions for CBM of written expression
What is Data-Based Individualization (DBI)?

- A decision-making framework for providing intensive academic intervention
  - Assumes good interventions don’t work for all students
- It generates evidence that either:
  - The intervention is working as designed
  - Your experimental teaching is working

National Center on Intensive Intervention, https://intensiveintervention.org/
Data-Based Individualization Requires Data

• Curriculum-based measurement (CBM) data are often used for this purpose
  • Indicators of overall progress in an academic skill area
  • Standardized
  • Efficient (easy to administer and score) and repeatable
  • Documented standards for performance
    • Criterion- and/or norm-referenced
  • Evidence of reliability and validity for screening and progress monitoring
    • Alternate-form reliability
    • Predicts performance on more comprehensive assessments of the skill
CBM Example: Oral Passage Reading

• Also called ‘oral reading fluency’
  • Read one or more field-tested passages for 1 min
  • Record the number of words read correctly

• Scores predict performance on comprehensive assessments of broad reading skill (Reschly et al., 2009)
  • Can identify students at-risk of difficulty/disability
  • Sensitive to improvements in general reading skill

• Easy to administer and use for decision making
  • Compare to norms
  • Graph data from repeated administrations and visually analyze progress
CBM in Written Expression (CBM-WE)

- The original idea (~1980s)
  - Present one narrative prompt (story starter: One day at school...)
  - 1 min to plan and 3 min to write
  - Score with simple metrics like word count

- This (and similar procedures) work pretty well in lower elementary grades for screening and monitoring, less so as student writing becomes more complex (McMaster & Espin, 2007)
  - Key issues: reliability, validity (including face validity), and feasibility
CBM-WE: Reliability

• Big Idea: Typical procedures do not yield reliable data for screening or progress monitoring

• Collected three 7 min narrative writing samples collected in fall, winter, and spring (n = 145 grade 2-5 students in Houston, TX, area)
  • Generalizability theory analyses to determine optimum sample duration and number of samples needed
  • Reliability < .80 for absolute screening decisions based on one 7 min sample
  • Reliability < .80 for decisions about student growth even with three 7 min writing samples

Keller-Margulis, Mercer, & Thomas (2016)
CBM-WE: Validity

• Big Idea: More complex scoring methods (than total words) improve validity, but greatly reduce feasibility

• Metrics like correct word sequences (CWS) have higher validity coefficients
  • Counts of the number of adjacent words that are spelled correctly and make sense in context
    • Considers aspects spelling, punctuation, syntax, and semantics
  • Better indicator of writing quality, but more time consuming and harder to reliably score
    • Feasibility concerns compound with multiple, longer duration writing samples
Potential Solution: Automated Text Evaluation

• Use computer software that considers and quantifies many characteristics of words, sentences, and discourse to evaluate CBM-WE writing samples
  • Commercial applications are already available, Project Essay Grade (Wilson, 2018)
    • It works well, but no info on how samples are scored and $$$
  • Develop open-source alternatives (Mercer et al., 2019)
    • Need to develop scoring models
    • Others can build on this work
    • Could be incorporated in data-management software
Current Project

• Can automated text evaluation be used to predict writing quality for longer duration narrative samples from students with substantial learning difficulties?
  • Convergent and discriminant validity (writing vs. reading and math)

• Are the scores sensitive to student skill growth from fall to spring?
Context and Sample

• Students participating in 1:1 academic intervention beyond school hours at the Learning Disability Society of Greater Vancouver (http://ldsociety.ca/)

• For training computer models:
  • 10 min picture-prompted narrative samples ($n = 204$) collected in Sep/Oct and May/June each year for program planning and evaluation from 105 students

• For evaluating validity:
  • Non-random sample of 33 students (grades 3-9) with standardized assessment scores in writing, reading, and math
Measures: Holistic Writing Quality

• Used to train automated text evaluation models for Sep/Oct and May/June picture-prompsted samples

• Paired comparison method (Thurstone, 1927)
  • Each rater identified best sample for 3000 pairs of samples
  • Aggregated to a continuous quality score using ranking algorithms
    • High inter-rater reliability ($r = .95$)

• Raters were asked to consider substantive quality (ideation, word choice, text structure)
  • Tiebreaker: Which sample would you most like to read more of?
Measures: Automated Writing Quality

- Each picture-prompted writing sample submitted to ReaderBench (Dascalu, Dessus, Trausan-Matu, Bianco, & Nardy, 2013)
  - Open-source software intended to assess text characteristics predicting reading comprehension difficulty
    - Provides ~200 indicators of word complexity, lexical diversity, syntactic complexity, cohesion, etc.

- Machine learning algorithms used to predict holistic quality ratings with RB scores as inputs
  - Partial least squares (PLS) regression worked best
  - 85% of variance in quality ratings explained
  - Algorithm-predicted quality used in validity analyses
Measures: Validity Assessments (May/June)

• Standardized Written Expression
  • Test of Written Language (4th ed.) constructed response (story writing)
    • Picture prompted, 5 min to plan, 15 min to write
    • Contextual Conventions (CC): spelling and grammar
    • Story Composition (SC): vocabulary, plot, interest to reader

• Standardized Broad Reading and Broad Math
  • aReading and aMath computerized adaptive tests
  • ~20 min to administer, assesses skills from K – Grade 12
  • https://charts.intensiveintervention.org/chart/academic-screening
# Results: Convergent and Discriminant Validity

Table 1. *Automated quality scores in relation to standardized writing, reading, and math scores*

<table>
<thead>
<tr>
<th></th>
<th>TOWL CC</th>
<th>TOWL SC</th>
<th>aReading</th>
<th>aMath</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r (R^2)$</td>
<td>$r (R^2)$</td>
<td>$r (R^2)$</td>
<td>$r (R^2)$</td>
</tr>
<tr>
<td>Fall Quality</td>
<td>.69 (.48)</td>
<td>.47 (.22)</td>
<td>.53 (.28)</td>
<td>.24 (.06)</td>
</tr>
<tr>
<td>Spring Quality</td>
<td>.76 (.57)</td>
<td>.53 (.28)</td>
<td>.56 (.31)</td>
<td>.35 (.12)</td>
</tr>
<tr>
<td>TOWL Quality</td>
<td>.78 (.60)</td>
<td>.69 (.48)</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

*Note.* $n = 33$. TOWL = Test of Written Language (4th ed.), CC = Contextual Conventions, SC = Story Composition. Values in italics are not statistically significant ($a = .05$).

Incremental validity compared to typical CBM-WE scoring

TWW: $r = .47$ and .59 with fall and spring TOWL CC; CWS: $r = .67$ and .67
Results: Sensitivity to Growth

• Statistically significant ($p < .001$), moderate-to-large overall change ($d = .77$) from fall to spring on automated quality scores
Discussion: Key Findings

• Good evidence of convergent and discriminant validity for use of automated text evaluation with agency-designed writing sample process to predict performance on more comprehensive assessments of academic skill
  • For students with significant learning difficulties participating in intensive intervention beyond school hours
  • Replicates and extends similar findings with a U.S. general education sample
  • Generalizability of automated scoring algorithm when applied to TOWL writing sample

• Automated quality scores showed evidence of student writing skill growth across a wide range of skill/grade levels (3-9)

• (Very) preliminary evidence that this could work for screening and progress monitoring in a DBI/CBM framework
Defensible Decisions Require Good Data

• Potentially very substantial improvements in scoring feasibility for screening and monitoring large numbers of students
  • Plus fewer concerns with inter-scorer agreement

• Can be used to generate local standards for performance (norms and criteria)
  • For identifying student needs, monitoring outcomes, evaluating programs, and allocating resources

• Not intended to replace evaluation of writing by teachers
  • Can assist teachers in evaluating and tracking overall quality, while freeing up time to provide detailed, formative feedback on areas to improve (Wilson & Czik, 2016)
Closing

• Acknowledgements
  • Funding
    • Social Sciences and Humanities Research Council of Canada
    • U.S. Department of Education, Institute of Education Sciences
      • The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.
    • Chris Spencer Foundation
  • Special Thanks
    • Staff and Students of the Learning Disabilities Society of Greater Vancouver

• More Information
  • Slides and paper: https://ecps.educ.ubc.ca/person/sterett-mercer/
References


