Searching more successfully for academic talent

Finding the right measures and using the right norm groups

David Lohman
Institute for Research & Policy on Acceleration
Overview

- A potpourri of recent research
  - Sex differences in Quantitative Reasoning
  - Stability of scores on ability & achievement tests
  - Frequency of extreme discrepancies in score profiles for the most (and least) able
- An aptitude *perspective* on talent
- Getting serious about Opportunity to Learn
- New developments
Sex differences in Quantitative Reasoning

Comparing Variability

- Sample size
- Ceiling effects
- Scaling & distributional assumptions
- Differential Item Functioning
  - Distinguish from sex differences
  - Magnified effects on extreme scores
Strand, Deary, & Smith (2006)

- Over 320,000 11-12 year old UK students
- Level D of the UK version of the Cognitive Abilities Test (CAT)
  - Verbal Battery (3 tests)
  - Quantitative Battery (3 tests)
  - Nonverbal Battery (3 tests)
Cognitive Abilities Test

Verbal
- Verbal Classification
- Verbal Analogies
- Sentence Completion

Quantitative
- Number Series
- Quantitative Relations
- Equation Building

Nonverbal
- Figure Analogies
- Figure Classification
- Figure Analysis

Number Analogies

3 Separate Test Batteries (not one)
Verbal Battery
Proportion M/F at each stanine
UK sample only

![Graph showing proportion of M/F at each stanine in UK sample only.](image)
Quantitative Battery - Level D
Proportion M/F at each stanine
UK sample only
Verbal Battery - Level D
Proportion M/F at each stanine
UK sample only

Proportion

stanine

UK M

UK F
Do these results generalize?

- To 6th grade students in the US?
- Across grades/test levels in 2000 US sample?
  - Developmental changes
  - Impact of a few items on extreme scores
U.S. Data
Conclusions

- Do results of Strand et al. (2006) generalize?
  - UK to US
  - Across grades (3 – 11)
  - Across assessments/cohorts

- Advantages of not changing the measure in a fundamental way
  - Versus NAEP & TIMMS

- Implications
Talent identification requires:

- A focus on potential rather than on current accomplishment
- Control for opportunity to learn
- The prediction of an unusual level future competence from present and past behavior
- Assumptions about the stability of individual differences in estimates of that competence
### Correlations Between IQ Scores across grades (Lohman & Korb, 2006)

<table>
<thead>
<tr>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>83</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>80</td>
<td>83</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>75</td>
<td>80</td>
<td>83</td>
<td>86</td>
</tr>
</tbody>
</table>
### Correlations Between IQ Scores across Grades (Lohman & Korb, 2006)

<table>
<thead>
<tr>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>83</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>80</td>
<td>83</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>75</td>
<td>80</td>
<td>83</td>
<td>86</td>
</tr>
</tbody>
</table>
### Proportion of students identified by one test also identified by the second test

<table>
<thead>
<tr>
<th>Cut score</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
<th>0.80</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1%</td>
<td>0.13</td>
<td>0.19</td>
<td>0.27</td>
<td>0.38</td>
<td>0.54</td>
</tr>
<tr>
<td>Top 2%</td>
<td>0.17</td>
<td>0.23</td>
<td>0.31</td>
<td>0.42</td>
<td>0.58</td>
</tr>
<tr>
<td>Top 3%</td>
<td>0.20</td>
<td>0.26</td>
<td>0.35</td>
<td>0.45</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Correlation between tests**
Individual differences will be stable to the extent that

- Individuals do not acquire new knowledge and skills at different rates.
- The opportunities that individuals have had to develop the measured competencies do not change.
- The aspects of competence that are assessed also do not change.

Implications?

- Assess cognitive development early and often
- Revolving door
Beyond a single rank order

- Brunswikian symmetry – picking the right predictors for the criteria of interest
N = 866

Ability Tests (Grade 6)

Test 1
Test 2
Test 3
Test 4
Test 5
Test 6
Test 7
Test 8
Test 9
Test 10
Test 11
Test 12
Test 13

Gustafsson & Balke, MBR, 1993
Gustafsson & Balke, MBR, 1993
Gustafsson & Balke, MBR, 1993

Ability Tests (Grade 6)
- Test 1
- Test 2
- Test 3
- Test 4
- Test 5
- Test 6
- Test 7
- Test 8
- Test 9
- Test 10
- Test 11
- Test 12
- Test 13

School Grades (Grade 9)
- Grade 1
- Grade 2
- Grade 3
- Grade 4
- Grade 5
- Grade 6
- Grade 7
- Grade 8
- Grade 9
- Grade 10
- Grade 11
- Grade 12
- Grade 13
- Grade 14
- Grade 15
- Grade 16
- Grade 17

V'
Ms'
N'
CFR'
Vz'
Cs'

Gc'

Lang'
SS'
Sci'
Pract'
Spatial'

.72
.75
.63
- g mattered most when predicting GPA
- Specific abilities mattered even more when predicting success in many domains
- Obtaining useful & reliable score profiles
Score Profiles on CogAT
Profiles for the most (and least) able

### CogAT6 Profile frequencies for students in K-12 pop. and for students with two stanine scores of 9

<table>
<thead>
<tr>
<th>Profile</th>
<th>Percent in K-12 population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>33</td>
</tr>
<tr>
<td>Significant (10 - 23)</td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>21</td>
</tr>
<tr>
<td>Weakness</td>
<td>22</td>
</tr>
<tr>
<td>Extreme (24 +)</td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>4</td>
</tr>
<tr>
<td>Weakness</td>
<td>3</td>
</tr>
</tbody>
</table>
### CogAT6 Profile frequencies for students in K-12 pop. and for students with two stanine scores of 9

<table>
<thead>
<tr>
<th>Profile</th>
<th>Percent in K-12 population</th>
<th>Percent in Stanine=9 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>33</td>
<td>37</td>
</tr>
<tr>
<td>Significant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Weakness</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Extreme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Weakness</td>
<td>3</td>
<td>16</td>
</tr>
</tbody>
</table>
Frequency of Flat Profiles

Median Stanine

Percent

Median Stanine

Percent
Profiles with a significant or extreme STRENGTH
Profiles with a significant or extreme WEAKNESS

![Graph showing the relationship between Median Stanine and Percent]
A brief introduction to an aptitude perspective on talent identification
Sources

- **Foundation in IO psychology**
  - Bingham (1937)

- **Dawes theory of work adjustment** (Lubinski)
  - Satisfactoriness versus Satisfaction

- **Aptitude/trait complexes** (Ackerman)
  - Constellation of traits that combine in a non-compensatory way

- **Gagne’s DMGT theory**
Talent identification and development

- Goal is to identify those who exhibit talent for learning in particular domains
- And then to provide appropriately challenging instruction to develop that talent
- Talent as a constellation of aptitudes
Aptitude: What it is

- Aptitude is the degree of *readiness* to learn & perform well in a *particular situation* or *domain*. 
Important Aptitudes for Academic Learning

- **Cognition (knowing)**
  - Domain knowledge & skill
  - Reasoning abilities in the symbol systems used to communicate knowledge (Verbal, Quant., Spatial)

- **Affection (feeling)**
  - anxiety, interests, working alone/with others

- **Conation (willing)**
  - persistence, impulsivity
Ways of measuring aptitude

1. Direct:
   - Learns in a few trials what takes others many trials to learn.
   - Opportunity to learn essential

2. Indirect:
   - Brings required (or helpful) cognitive, affective, & conative resources to the situation
   - Opportunity to learn essential
Opportunity to Learn

- On ability tests:
  - Approximate by age (years + months)

- On achievement tests
  - Approximate by grade (and week within grade)

- What about students who are poor, ELL, or have markedly different educational preparation?
Non-normative Experiences
Non-normative Experiences

Alternative assessments (common norms)
Advantages of nonverbal tests

- English Language Learners are disadvantaged on tests that use English
  - Allow one to use common norms for all
- Often good measures of g
- Appearance of measuring something more innate than tests that use other symbols
- Claims that they will solve the problem
Disadvantages of nonverbal tests

- Construct under-representation
- Predict academic learning less well than measures of verbal & quant reasoning
  - Specific variance negatively related to success in verbal domains
- Sensitive to environmental changes:
  - Large Flynn, practice effects, and schooling effects
- (P.S. Not tests of spatial ability)
Level the playing field?

- 371 ELL & 332 Non-ELL K-6 children
  - All Hispanic
  - All free/reduced lunch
- ELL – Non-ELL score differences
  - CogAT Nonverbal 7.5 (.47 SD)
  - Raven SPM 7.3 (.46 SD)
  - NNAT 10.1 (.63 SD)
Trusting the Norms

- Judgments of exceptionality depend critically on the appropriateness & quality of the norms
- Raven Norms – about 10 points too high
  - 100.7 versus 111.5 (n = 733)
- Cattell Culture Fair – about 17 points too high
- CogAT Quantitative Battery
  - Shift in educational practice
- NNAT?
SD's of Nonverbal Ability Index (NAI) scores on NNAT by Test Level

Standard Deviation

NNAT Test Level

A B C D E F G

Expected
SD's of Nonverbal Ability Index (NAI) scores on NNAT by Test Level

![Graph showing SD's of NAI scores on NNAT by Test Level]

- Expected
- Bright Horizon
SD's of Nonverbal Ability Index (NAI) scores on NNAT by Test Level

Standard Deviation

NNAT Test Level

Expected
Bright Horizon
NNAT Spring
SD's of Nonverbal Ability Index (NAI) scores on NNAT by Test Level

- Expected
- Bright Horizon
- NNAT Spring
- NNAT Fall

NNAT Test Level

Standard Deviation

0 5 10 15 20 25

A B C D E F G
# Over-identification Rates for NNAT

<table>
<thead>
<tr>
<th>Level</th>
<th>True NAI Score 115</th>
<th>True NAI Score 130</th>
<th>True NAI Score 145</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.5</td>
<td>3.4</td>
<td>11.9</td>
</tr>
<tr>
<td>B</td>
<td>1.4</td>
<td>2.6</td>
<td>7.3</td>
</tr>
<tr>
<td>C</td>
<td>1.3</td>
<td>2.3</td>
<td>5.8</td>
</tr>
<tr>
<td>D</td>
<td>1.2</td>
<td>1.7</td>
<td>2.9</td>
</tr>
<tr>
<td>E</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>1.1</td>
<td>1.4</td>
<td>2.0</td>
</tr>
<tr>
<td>G</td>
<td>1.1</td>
<td>1.4</td>
<td>1.9</td>
</tr>
</tbody>
</table>
## Over-identification Rates for NNAT

<table>
<thead>
<tr>
<th>Level</th>
<th>True NAI Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>115</td>
</tr>
<tr>
<td>A</td>
<td>1.5</td>
</tr>
<tr>
<td>B</td>
<td>1.4</td>
</tr>
<tr>
<td>C</td>
<td>1.3</td>
</tr>
<tr>
<td>D</td>
<td>1.2</td>
</tr>
<tr>
<td>E</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>1.1</td>
</tr>
<tr>
<td>G</td>
<td>1.1</td>
</tr>
<tr>
<td>Level</td>
<td>True NAI Score</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>115</td>
</tr>
<tr>
<td>A</td>
<td>1.5</td>
</tr>
<tr>
<td>B</td>
<td>1.4</td>
</tr>
<tr>
<td>C</td>
<td>1.3</td>
</tr>
<tr>
<td>D</td>
<td>1.2</td>
</tr>
<tr>
<td>E</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>1.1</td>
</tr>
<tr>
<td>G</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Over-identification Rates for NNAT

Caveat emptor

- Critically examine your data
- If possible, always use local norms as well
Means versus correlations

On the whole, academic success for ELL and minority children requires the same knowledge, skills, abilities, interests, motivation, perseverance as their non-minority peers.
White-Hispanic Effect Sizes for CogAT (random samples within bldg)

Nonverbal the fairest test?
## Project Bright Horizon Data

### ELL Students Grade 3 (N = 74 - 81)

<table>
<thead>
<tr>
<th>Test</th>
<th>CogAT V</th>
<th>CogAT Q</th>
<th>CogAT N</th>
<th>NNAT</th>
<th>Raven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev. Reading Assess.</td>
<td>0.63</td>
<td>0.43</td>
<td>0.24</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>Reading - Terra Nova</td>
<td>0.56</td>
<td>0.36</td>
<td>0.30</td>
<td>0.33</td>
<td>0.41</td>
</tr>
<tr>
<td>Reading - AIMS</td>
<td>0.73</td>
<td>0.45</td>
<td>0.34</td>
<td>0.39</td>
<td>0.43</td>
</tr>
<tr>
<td>Math - Terra Nova</td>
<td>0.57</td>
<td>0.61</td>
<td>0.51</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>Math - AIMS</td>
<td>0.61</td>
<td>0.70</td>
<td>0.57</td>
<td>0.50</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Is the problem the test?

Or the norm group?
Non-normative Experiences

Alternative assessments (common norms)

Common assessments (multiple norm groups)
Multiple Perspectives

- For ELL students in grade 3, compare scores to:
  - Other grade 3 students in the nation (common norms)
  - Other students in grade 3 in the district/school
  - Other ELL students in grade 3 in the district

- Not a single statistical adjustment of scores (Mercer)

- Not an assertion that all are ready for the same level of instruction
Tradeoff

Measuring the right things approximately for ELL students

or

the wrong things with greater accuracy
Improvements in Measurement for ELL children

1. One test for all (e.g., early Binet)
2. “Nonverbal” tests (Army Beta)
   Reduced difference between mean scores of ELL and native speakers of English
3. Translated/adapted English-language tests (Woodcock-Munoz)
4. Translated tests with contextualized norms (WISC-IV Spanish)
5. Simultaneous development in (both) languages with contextualized norms (Current work)
WISC-IV Spanish

- Target population: Bilingual Spanish-speaking children in the U.S. with no more than 5 years of U.S. schooling
- Translated & adapted the verbal tests
- Calibrated the Spanish Verbal Scale with the normative (English) Verbal Scale
- Indexed opportunity to learn by
  - % child’s education in U.S.
  - Parental educational level
WISC-IV Spanish Scores by Percent Education in the U.S.

Scale Score (M = 100, SD = 15)

- Minimal (<30%)
- Most
- All (100%)

WISC-IV Score

Full Scale IQ  Verbal Comp  Perceptual Reasoning
WISC-IV Spanish Scores by Percent Education in the U.S.

Scale Score (M = 100, SD = 15)

- Minimal (<30%)
- Most
- All (100%)

WISC-IV Score

- Full Scale IQ
- Verbal Comp
- Perceptual Reasoning
WISC-IV Spanish Scores by Percent Education in the U.S.

Scale Score (M = 100, SD = 15)

- Minimal (<30%)
- Most
- All (100%)

WISC-IV Score

Full Scale IQ  Verbal Comp  Perceptual Reasoning
“nonlanguage tests may be more culturally loaded than language tests”

Anastasi & Urbina, 1997, p. 344
WISC-IV Spanish Scores: Multiple Perspectives

- **Index scores (Mean 100, SD 15)**
  - Verbal Comprehension (Spanish Calibration)
  - Perceptual Reasoning
  - Working memory
  - Processing Speed

- **PR (by Pct Education in U.S.)**

- **PR (by Pct Education in U.S & Parental Education)**

- Professional judgment required
Main Points

- Focus on talent identification & development
- Instability of observed scores on the best tests
- Importance of affective & conative traits
- Assumptions about opportunity to learn
- Consider alternative norms rather than alternative assessments
- Develop tests for bilingual children that are not simply translated versions of the English language test
Thank you.

faculty.education.uiowa.edu/dlohman