Smarter Balanced Scores – Guidance for Different Users

Presented to California Educational Research Association

December 3, 2015

Joe Willhoft, Assessment Consultant
Deb Sigman, Deputy Director Standards, Assessment, and Accountability Services
Topics for Discussion

1. Smarter Balanced Vertical Scale
2. Assessment Precision for Individuals and Groups
3. Connecting Assessment Purpose to Use
4. Using Results for District-Level Planning
5. Aligning Policies with Best Assessment Practices
Smarter Balanced Vertical Scales
Things to Know About Smarter Balanced

1. An adaptive test
   - Test questions are individualized
   - The test questions vary from student to student
An 11-item adaptive test
How is adaptive different from a Fixed Form test?

- **Fixed**: Forms are common for all students
- **Adaptive**: Test “forms” are individualized

- **Fixed**: Students are “grouped” into scale score intervals (Not all scale scores are possible)
- **Adaptive**: Each student receives his/her “earned” score (All scale scores are possible)
## CST Raw Score to Scale Score ELA – Grade 7

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Scale Score</th>
<th>Reported Score</th>
<th>Raw Score</th>
<th>Scale Score</th>
<th>Reported Score</th>
<th>Raw Score</th>
<th>Scale Score</th>
<th>Reported Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0006</td>
<td>150</td>
<td>26</td>
<td>264.1642</td>
<td>264</td>
<td>51</td>
<td>354.0178</td>
<td>354</td>
</tr>
<tr>
<td>1</td>
<td>43.7005</td>
<td>150</td>
<td>27</td>
<td>267.7729</td>
<td>268</td>
<td>52</td>
<td>357.8742</td>
<td>358</td>
</tr>
<tr>
<td>2</td>
<td>85.0430</td>
<td>150</td>
<td>28</td>
<td>271.3679</td>
<td>271</td>
<td>53</td>
<td>361.7826</td>
<td>362</td>
</tr>
<tr>
<td>3</td>
<td>109.7695</td>
<td>150</td>
<td>29</td>
<td>274.9535</td>
<td>275</td>
<td>54</td>
<td>365.9938</td>
<td>366</td>
</tr>
<tr>
<td>4</td>
<td>127.4037</td>
<td>150</td>
<td>30</td>
<td>278.5287</td>
<td>279</td>
<td>55</td>
<td>370.3851</td>
<td>370</td>
</tr>
<tr>
<td>5</td>
<td>141.2302</td>
<td>150</td>
<td>31</td>
<td>282.0941</td>
<td>282</td>
<td>56</td>
<td>374.7425</td>
<td>375</td>
</tr>
<tr>
<td>6</td>
<td>152.7589</td>
<td>155</td>
<td>32</td>
<td>285.6490</td>
<td>286</td>
<td>57</td>
<td>379.2740</td>
<td>379</td>
</tr>
<tr>
<td>7</td>
<td>162.8320</td>
<td>163</td>
<td>33</td>
<td>289.1930</td>
<td>289</td>
<td>58</td>
<td>384.0714</td>
<td>384</td>
</tr>
<tr>
<td>8</td>
<td>171.6637</td>
<td>172</td>
<td>34</td>
<td>292.7250</td>
<td>293</td>
<td>59</td>
<td>388.9130</td>
<td>389</td>
</tr>
<tr>
<td>9</td>
<td>179.6233</td>
<td>180</td>
<td>35</td>
<td>296.2450</td>
<td>296</td>
<td>60</td>
<td>394.0079</td>
<td>394</td>
</tr>
<tr>
<td>10</td>
<td>186.8424</td>
<td>187</td>
<td>36</td>
<td>299.7527</td>
<td>300</td>
<td>61</td>
<td>399.3018</td>
<td>399</td>
</tr>
<tr>
<td>11</td>
<td>193.5681</td>
<td>194</td>
<td>37</td>
<td>303.2465</td>
<td>303</td>
<td>62</td>
<td>404.9254</td>
<td>405</td>
</tr>
<tr>
<td>12</td>
<td>199.9678</td>
<td>200</td>
<td>38</td>
<td>306.7265</td>
<td>307</td>
<td>63</td>
<td>410.6998</td>
<td>411</td>
</tr>
<tr>
<td>13</td>
<td>205.7389</td>
<td>206</td>
<td>39</td>
<td>310.1918</td>
<td>310</td>
<td>64</td>
<td>416.7990</td>
<td>417</td>
</tr>
<tr>
<td>14</td>
<td>211.2240</td>
<td>211</td>
<td>40</td>
<td>313.6420</td>
<td>314</td>
<td>65</td>
<td>423.7611</td>
<td>424</td>
</tr>
<tr>
<td>15</td>
<td>216.5488</td>
<td>217</td>
<td>41</td>
<td>317.0763</td>
<td>317</td>
<td>66</td>
<td>431.0093</td>
<td>431</td>
</tr>
<tr>
<td>16</td>
<td>221.6289</td>
<td>222</td>
<td>42</td>
<td>320.5646</td>
<td>321</td>
<td>67</td>
<td>438.9023</td>
<td>439</td>
</tr>
<tr>
<td>17</td>
<td>226.4189</td>
<td>226</td>
<td>43</td>
<td>324.3560</td>
<td>324</td>
<td>68</td>
<td>447.7370</td>
<td>448</td>
</tr>
<tr>
<td>18</td>
<td>231.2920</td>
<td>231</td>
<td>44</td>
<td>327.8525</td>
<td>328</td>
<td>69</td>
<td>457.7945</td>
<td>458</td>
</tr>
<tr>
<td>19</td>
<td>235.6908</td>
<td>236</td>
<td>45</td>
<td>331.3934</td>
<td>331</td>
<td>70</td>
<td>469.5002</td>
<td>470</td>
</tr>
<tr>
<td>20</td>
<td>240.0747</td>
<td>240</td>
<td>46</td>
<td>335.1744</td>
<td>335</td>
<td>71</td>
<td>483.7134</td>
<td>484</td>
</tr>
<tr>
<td>21</td>
<td>244.3656</td>
<td>244</td>
<td>47</td>
<td>338.7738</td>
<td>339</td>
<td>72</td>
<td>501.7728</td>
<td>502</td>
</tr>
<tr>
<td>22</td>
<td>248.2563</td>
<td>248</td>
<td>48</td>
<td>342.4324</td>
<td>342</td>
<td>73</td>
<td>526.9643</td>
<td>527</td>
</tr>
<tr>
<td>23</td>
<td>252.2015</td>
<td>252</td>
<td>49</td>
<td>346.2195</td>
<td>346</td>
<td>74</td>
<td>565.8094</td>
<td>566</td>
</tr>
<tr>
<td>24</td>
<td>256.3791</td>
<td>256</td>
<td>50</td>
<td>350.1320</td>
<td>350</td>
<td>75</td>
<td>625.1589</td>
<td>600</td>
</tr>
<tr>
<td>25</td>
<td>260.2167</td>
<td>260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Things to Know About Smarter Balanced (cont.)

1. An adaptive test

2. Content Rules!
   - Though the test questions vary from student to student, the CCSS content is the same for all students
   - SB uses common “test blueprints” for each grade level/content area
Two 11-item adaptive tests
Things to Know About Smarter Balanced (cont.)

1. An adaptive test

2. Content Rules!

3. Scores reported on an across-grade “Vertical Scale”
   - Vertical Scales can be used to judge growth over time
   - Scale uses scores in the 2000-3000 range
Vertical Scale English/Language Arts
(Field test dataset)
Vertical Scale for Mathematics (Field test dataset)
Things to Know About Smarter Balanced (cont.)

1. An adaptive test
2. Content Rules!
3. Scores reported on an across grade Vertical Scale

4. Threshold ("cut") scores also on the Vertical Scale
   - Helps in understanding of achievement levels
   - Useful for goal-setting
<table>
<thead>
<tr>
<th>Grade</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2367</td>
<td>2432</td>
<td>2490</td>
</tr>
<tr>
<td>4</td>
<td>2416</td>
<td>2473</td>
<td>2533</td>
</tr>
<tr>
<td>5</td>
<td>2442</td>
<td>2502</td>
<td>2582</td>
</tr>
<tr>
<td>6</td>
<td>2457</td>
<td>2531</td>
<td>2618</td>
</tr>
<tr>
<td>7</td>
<td>2479</td>
<td>2552</td>
<td>2649</td>
</tr>
<tr>
<td>8</td>
<td>2487</td>
<td>2567</td>
<td>2668</td>
</tr>
<tr>
<td>11</td>
<td>2493</td>
<td>2583</td>
<td>2682</td>
</tr>
</tbody>
</table>
## Smarter Balanced Mathematics

### Threshold (Cut) Scores

<table>
<thead>
<tr>
<th>Grade</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2381</td>
<td>2436</td>
<td>2501</td>
</tr>
<tr>
<td>4</td>
<td>2411</td>
<td>2485</td>
<td>2549</td>
</tr>
<tr>
<td>5</td>
<td>2455</td>
<td>2528</td>
<td>2579</td>
</tr>
<tr>
<td>6</td>
<td>2473</td>
<td>2552</td>
<td>2610</td>
</tr>
<tr>
<td>7</td>
<td>2484</td>
<td>2567</td>
<td>2635</td>
</tr>
<tr>
<td>8</td>
<td>2504</td>
<td>2586</td>
<td>2653</td>
</tr>
<tr>
<td>11</td>
<td>2543</td>
<td>2628</td>
<td>2718</td>
</tr>
</tbody>
</table>
English Language Arts/Literacy Threshold Scale Scores

Threshold Scale Score vs. Grade

- Level 4
- Level 3
- Level 2

Grade: 3, 4, 5, 6, 7, 8, 9, 10, 11
Mathematics
Threshold Scale Scores

Threshold Scale Score vs Grade

- Level 4
- Level 3
- Level 2

Grade: 3 4 5 6 7 8 9 10 11

WestEd
Things to Know About Smarter Balanced – Recap

1. An adaptive test
2. Content Rules!
3. Scores reported on an across grade Vertical Scale
4. Threshold (“cut”) scores also on the Vertical Scale
Assessment Precision
How Precise Is A Score?

Stephen Curry

31.6 points/game thru 11/30

How many points will he score tonight?
1. For most tests, the standard error is about 1/3 of a standard deviation
   - Usually called the “SEM” (standard error of measurement)
   - On a “Fixed Form” test, this is true for the middle of the distribution
   - On an adaptive test, this holds for more students at the extremes
# Means and Standard Deviations for ELA and Math

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Scale Scores</th>
<th>Observed Range</th>
<th>Field test data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>(Std. Dev.)</td>
<td>Low</td>
</tr>
<tr>
<td>ELA</td>
<td>3</td>
<td>2401.9</td>
<td>(91.0)</td>
<td>2114</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2444.1</td>
<td>(95.4)</td>
<td>2131</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2481.6</td>
<td>(94.2)</td>
<td>2201</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2503.5</td>
<td>(94.8)</td>
<td>2210</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2518.0</td>
<td>(96.9)</td>
<td>2259</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2541.0</td>
<td>(96.8)</td>
<td>2288</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2554.3</td>
<td>(103.1)</td>
<td>2299</td>
</tr>
<tr>
<td>MATH</td>
<td>3</td>
<td>2413.0</td>
<td>(77.5)</td>
<td>2189</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2458.8</td>
<td>(79.3)</td>
<td>2204</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2487.6</td>
<td>(86.2)</td>
<td>2219</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2507.0</td>
<td>(94.9)</td>
<td>2235</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2515.8</td>
<td>(106.0)</td>
<td>2250</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2528.9</td>
<td>(112.7)</td>
<td>2265</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2568.0</td>
<td>(125.0)</td>
<td>2280</td>
</tr>
</tbody>
</table>
User Tips for Standard Errors (cont.)

1. For most tests, the SEM is about 1/3 of a standard deviation.

2. A good working estimate of the SEM is about 30 scale score points:
   - For a student, add and subtract about 30 points to get a range of “likely” scores.
   - So, a student with a score of 2550 would likely score between 2520-2580 if tested again.
User Tips for Standard Errors (cont.)

1. For most tests, the SEM is about 1/3 of a std. deviation

2. A good working estimate of the SEM is about 30 scale score points

3. The “30-6-3-1 Rule”
   - For 1 student, SEM is about 30 points
   - For a “classroom” (~25), SEM is about 6 points
   - For a “grade team” (~100) SEM is about 3 points
   - For a “district” (>1,000) SEM is about 1 point
User Tips for Standard Errors (Cont.)

1. For most tests, the SEM is about 1/3 of a std. deviation
2. A good working estimate of the SEM is about 30 scale score points
3. The “30-6-3-1 Rule”
4. SB “Claim Scores” use the SEM around the Level 2/3 Cut Score as a “Borderline” category
Claim Scores

- Claim Scores (e.g., “Reading”) indicate if student is:
  - “Below Standard”
    - Clearly below the Level 2/3 cut score;
  - “Above Standard”
    - Clearly above the Level 2/3 cut; or
  - “At or Near Standard”
    - Too close to call
Connecting Assessment Purpose to Use
Getting it Right

• “Too often, testing is treated narrowly, rather than as a flexible tool to obtain information about important questions.”

Right Purpose – Right Use

- Think about why we have an outcry about “too much testing”

“... it is not clear that some of the tests that school districts administer were designed for the purposes for which they are used. The most controversial example is the use of state summative exams to evaluate school district staff when most of these tests were designed to track district and school progress, not individual staff-member proficiency.”

Student Testing in America’s Great City Schools: An Inventory and Preliminary Analysis, October 2015
Making the Most of Scores

Summative Assessment

Teaching & learning
- Responding to feedback
- Assessment

Feedback to learners & teaching process
- Analysis of results

WestEd
A Balanced Assessment System

Standards set expectations on path to college and career readiness

Summative assessments benchmarked to college and career readiness (Grades 3-8 and 11)

Teachers and schools have information and tools to improve teaching and learning

Interim assessments flexible, open, used for actionable feedback

All students graduate college and career ready

Digital Library: Formative assessment tools and practices for teachers to improve instruction
Results of a Statewide Summative Assessment Denote the Tip of the Iceberg – It Pays to Pay Attention
Making the Most of Summative Results

- Data (scores) are necessary, but not sufficient
  - Be aware of data stall
- Focus on groups, programs and disaggregation (not individuals)
- Always consider the culture, climate and context of your school/district.
- Provide a comparable context
- Will be different depending on audience
  - Parents
  - Teachers
  - Students
  - Local Board
  - Broader Community
Making the Most of Summative Results (cont.)

- Rarely provide definitive answers, but raise many questions allowing reflection on practice
- Provide a general information/direction ad a necessary story
- Provide an entry point into a collaborative, honest conversation
- Not the whole story, but a necessary story
- Must dig deeper to determine cause
Data Examination

• Comprehensive, complex, difficult process
  – Not a checklist
  – Not meant to be completed at a single staff meeting or a single PD event

• **Focus must be on improving learning**
  – Not solely about increasing scores

• Collaborative process that requires
  – Honesty
  – Willingness to commit the time
  – Ability to handle ambiguity
  – Patience
Data Examination (cont.)

• There are multiple ways to look at data
  – Maximize data pictures and perspectives
  – Different displays can elicit different responses

• Every district/site has its own culture/climate/context –
  – Examine data within the particular context of district/site to emphasize particular point or encourage a particular conversation

• Determine what you want audience to walk away with
  – What’s the next conversation?
Data Examination (cont.)

- Use the tools available to manage the data.
- Use tools effectively to reflect on instructional practice and standards implementation.
- Use data to inform practice and improve teaching and learning.
Using the Results for District-Level Planning
Importance of Assessment Data, Local Control Funding Formula and LCAP

In June of 2013, a new era of school finance in California was signed into law. This new funding model, LCFF, reshapes school funding, and is squarely aimed to improve achievement for all students.

LCFF, and its local accountability counterpart, the LCAP, are anchored by the notion that California must do better for its underperforming students.
California ELA/Literacy Achievement Disparities:
Percentage of Students Meeting or Exceeding Standards
California Math Achievement Disparities:
Percentage of Students Meeting or Exceeding Standards
Call to Action - Pay Attention but Move Beyond the Scores

• Comprehensive, complex, difficult process
  - Not a check list
  - Not meant to be completed at a single staff meeting or a single PD event

• Focus on improving learning
  - Not solely about increasing scores
  - Data not to be dismissed
  - Use baseline as a call to action to move forward
Call to Action - Pay Attention but Move Beyond the Scores

• Data (scores) are necessary, but not sufficient
• Reflect on what you can control to move beyond the scores:
  — Policies and Procedures
  — Practice
  — People
  — Programs
Aligning Policies with Best Assessment Practices – Why Does It Matter?
Aligning Policies with Practice

- Standards for Educational and Psychological Testing
  - Use and interpretation of Educational Assessment

- Examples
  - Algebra requirement
  - Using single assessments to promote/retain
GOOD INTENTIONS
BAD RESULTS