Validity and Reliability of a New Pediatric-Specific Multisource Feedback Evaluation Tool (Peds360) for Residents

Su-Ting T. Li, MD, MPH¹
Jamal Abedi, PhD¹
Daniel C. West, MD²

¹UC Davis
²UC San Francisco
Disclosures

- The authors have no potential conflicts of interest
Background

- Multisource feedback evaluations
  - Required by Pediatric Review Committee
  - Can positively affect pediatric resident professionalism and communication skills

Background

- Problems with existing multisource feedback evaluation tools
  - ABIM tool developed to assess adult patient encounters with practicing physicians
    - Not adequately designed and tested to assess pediatric encounters with pediatric residents
  - Other published tools
    - Designed to assess encounters with practicing pediatricians
    - Tools designed to assess pediatric residents not previously examined for validity and reliability
  - Critical problem: Lack of information about how to administer for high stakes decision making


Objectives

- Develop a new pediatric-specific multisource feedback evaluation tool (Peds360)
- Establish validity and reliability of Peds360
- Determine the best way to administer the tool for high stakes decision making (i.e. making decision about competency)
Peds360 Tool Design

- Modified ABIM’s patient survey to create pediatric-specific surveys
  - Patient/Family (Pt)
  - Health care professional (HCP)
  - Peer
  - Self

- Altered response scale
  - Poor Fair Good Very good Excellent
  - SD Disagree Neither Agree SA
Methods

- Used Peds360 to evaluate 36 pediatric residents over 1 yr

- Validity
  - Structural equation modeling

- Reliability
  - Generalizability-study: 2 different, 2-facet studies
    - Occasion * Domain (Prof, ICS, PC)
    - Within each Domain: Occasion * Item
  - Decision-study (for G-coefficient >0.8)
    - Estimate number of evaluations needed for good reliability
## Results

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Number of surveys collected</th>
<th>Total</th>
<th>Mean (#/resident)</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pt</td>
<td>932</td>
<td>25.9</td>
<td>9-47</td>
<td></td>
</tr>
<tr>
<td>HCP</td>
<td>1881</td>
<td>52.3</td>
<td>14-113</td>
<td></td>
</tr>
<tr>
<td>Peer</td>
<td>265</td>
<td>7.4</td>
<td>3-15</td>
<td></td>
</tr>
</tbody>
</table>
## High Construct Validity

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pt</td>
<td>0.87</td>
</tr>
<tr>
<td>HCP</td>
<td>0.90</td>
</tr>
<tr>
<td>Peer</td>
<td>0.92</td>
</tr>
</tbody>
</table>
G-study:
Sources of variance for composite score

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>O</th>
<th>D</th>
<th>R*O</th>
<th>R*D</th>
<th>O*D</th>
<th>R<em>O</em>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pt</td>
<td>0.1</td>
<td>0.6</td>
<td>78.7</td>
<td>12.8</td>
<td>&lt;0.1</td>
<td>0.3</td>
<td>7.4</td>
</tr>
<tr>
<td>HCP</td>
<td>6.7</td>
<td>0.5</td>
<td>62.1</td>
<td>16.0</td>
<td>0.8</td>
<td>0.1</td>
<td>13.8</td>
</tr>
<tr>
<td>Peer</td>
<td>49.2</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>19.5</td>
<td>0.2</td>
<td>1.1</td>
<td>30.0</td>
</tr>
</tbody>
</table>

R = resident (subject); O = occasion; D = domain
# G-study: Sources of variance by domain score

<table>
<thead>
<tr>
<th>Domain</th>
<th>R</th>
<th>O</th>
<th>I</th>
<th>R*O</th>
<th>R*I</th>
<th>O*I</th>
<th>R<em>O</em>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pt - Prof</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>7.8</td>
<td>33.2</td>
<td>&lt;0.1</td>
<td>0.5</td>
<td>58.5</td>
</tr>
<tr>
<td>Pt - ICS</td>
<td>1.3</td>
<td>1.9</td>
<td>2.0</td>
<td>36.4</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>58.3</td>
</tr>
<tr>
<td>Pt - PC</td>
<td>&lt;0.1</td>
<td>4.7</td>
<td>0.8</td>
<td>40.3</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>54.2</td>
</tr>
<tr>
<td>HCP - Prof</td>
<td>9.4</td>
<td>0.5</td>
<td>3.0</td>
<td>25.3</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>61.8</td>
</tr>
<tr>
<td>HCP - ICS</td>
<td>10.5</td>
<td>0.6</td>
<td>2.6</td>
<td>22.9</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>63.3</td>
</tr>
<tr>
<td>HCP - PC</td>
<td>9.7</td>
<td>1.1</td>
<td>2.2</td>
<td>17.5</td>
<td>1.1</td>
<td>0.7</td>
<td>67.8</td>
</tr>
<tr>
<td>Peer - Prof</td>
<td>36.5</td>
<td>0.1</td>
<td>&lt;0.1</td>
<td>10.8</td>
<td>4.3</td>
<td>&lt;0.1</td>
<td>48.3</td>
</tr>
<tr>
<td>Peer - ICS</td>
<td>19.7</td>
<td>0.1</td>
<td>1.5</td>
<td>17.6</td>
<td>5.5</td>
<td>&lt;0.1</td>
<td>55.6</td>
</tr>
</tbody>
</table>

R = resident; O = occasion; I = survey item
## D-study results

<table>
<thead>
<tr>
<th></th>
<th>Composite score</th>
<th>Prof</th>
<th>ICS</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pt</strong></td>
<td>440</td>
<td>&gt;1000</td>
<td>180</td>
<td>&gt;1000</td>
</tr>
<tr>
<td><strong>HCP</strong></td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td><strong>Peer</strong></td>
<td>3</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Construct validity high
- Reliability varies greatly depending on source (pt, HCP, peer) and domain assessed
  - HCP and peer evaluations were reliable, requiring a feasible number of evaluations
  - Patient evaluations were unreliable
- Professionalism domain reliability problems may reflect the complexity of that construct
Limitations

- Single institution
- Reverse-worded questions ("Talked down to me.") were eliminated because of too much variability in response indicating possible carelessness
- Peer evaluations may not reflect true performance (too nice)

Next steps

- Revise Peds360 to include non-reverse-worded questions (“Did not talk down to me.”)
- Re-examine validity and reliability
- Revise professionalism to better capture complexity of construct
- Revise interpersonal and communication domains to improve face validity for trainees and evaluators (HCP, peer)
- Examine external validity by testing tools in other institutions