A Systematic Review
Establishing Absolute Standards for Technical Performance

Authors: Mitchell G. Goldenberg, Alaina Garbens, Peter Szasz, Tyler Hauer, Teodor P. Grantcharov
Date: October 1st, 2016
I have/had an affiliation (financial or otherwise) with a pharmaceutical, medical device or communications organization.

J’ai (ou j’ai eu) une affiliation (financière ou autre) avec une entreprise pharmaceutique, un fabricant d’appareils médicaux ou un cabinet de communication.

Dr. Grantcharov (PI) has received research grants from:
Ethicon, Intuitive, Takeda, Covidien, Olympus

No other authors have conflicts to declare
Outline

• This presentation will address the following questions
  1. Why do we need standards in technical performance?
  2. What’s in a standard? What makes one *absolute*?
  3. What has been *achieved to date*?
  4. What is the *quality* of the evidence?
  5. How *have we/can we* best establish absolute standards in procedural performance going forward?
Why Do We Need Procedural Standards?
Educational Models in Procedural Assessment

Fitts and Posner (1967) ➔ Advanced Beginner ➔ Competent

Dreyfus and Dreyfus (1986) ➔ Knows How ➔ Shows How

Miller’s Pyramid (1990)
Why Do We Need Procedural Standards?

• Competency-Based Medical Education
  ➢ Formative Assessment
    » Directed feedback at low-stakes assessments
    » **Emphasis on educational process and feedback**
  ➢ Summative Assessment
    » Pass/Fail Decisions at high-stakes assessments
    » **Emphasis on outcome, defensibility**

• In Accreditation
  ➢ Incorporation into College/Board examinations
What is Standard Setting?

• Creation of ‘Cutoff’ point in an assessment

• *Relative Standards*
  - Standard reflects the **performance of another, well-defined group**
    - Eg. Pass mark = 75<sup>th</sup> percentile, 2 SD below the expert mean
    - Standard ‘dependent’ on index group performance
  - Simple to create/use, but not based on expert consensus

• *Absolute Standards*
  - Standard reflects the judgment/consensus of **expert ‘judges’**
    - Eg. Pass mark = 70%, 18/20
  - Involves conceptualization of ‘borderline’ student/participant
Absolute Standards

• *Item-Centred vs. Participant-Centred*

  ➤ **Item-Centred** (e.g. Angoff, Ebel, Hofstee)
    » Judgments are pre-formed, based on test content/difficulty
    » Suitable when ‘non-test factors’ can be controlled for
      • i.e. VR, Simulation
    » Based on the conceptualization of the ‘borderline student’

  ➤ **Participant-Centred** (e.g. Contrasting Groups, Borderline Group)
    » Judgments made on observations, based on individual performance
    » Suitable when expert judges able to directly observe performances
      • i.e. OSCE, Work-Based Assessments
Item-Centred (Example)

- **Angoff Method**
  - Judges reach consensus on ‘borderline’ student characteristics
  - Judges individually estimate borderline student’s performance on each test item (0% to 100%, or ‘yes/no’)
  - Judgments are systematically combined to determine pass mark

![Table 1. Sample Angoff Ratings and Calculation of Angoff Passing Score](image)
Participant-Centred (Example)

- **Contrasting-Groups**
  - Judges rate performances using checklist or rating scale
  - Application of ‘competent/non-competent’ grouping based on either
    - External criteria (i.e. trainee vs. staff)
    - Global pass/fail decision
  - Scores distributed on graph in groupings
  - Pass score = intersection of curves
How has absolute standard setting been used in procedural assessment?
Systematic Review Methodology

- PRISMA Protocol
- MEDLINE, Embase, PsychInfo
- Included:
  - Medical/Surgical/OBGYN
  - All environments
    - Simulation, Bedside, Operating Room, etc.
  - All levels of training (medical students → staff physicians)
- Excluded:
  - Non-procedural skills
    - i.e. CPR, acute-care management
- Quality of evidence assessed using the Medical Education Research Quality Instrument (MERSQI)
Results

Literature Identified Through Database Search, n=1809

Articles Excluded After Screening Titles and Abstracts, n=1638

Studies Identified Through Bibliography Searches, n=11

Total Full Text Review Articles, n=182

Articles Included for Full Text Review, n=171

Articles Included in Review, n=37

Articles Excluded in Full Text Review, n=135
- Arbitrarily chosen pass mark (14)
- Non-procedural assessment (31)
- Pass/fail decision made a priori (34)
- Relative Standard (10)
- Previously established pass mark (19)
- Pre-post assessment (3)
- No standard set (24)
- Not original research (i.e. review article) (8)
- Duplicate (2)
Participants

- 2 (5%) included *medical students*
- 26 (70%) included *interns/residents*
- 6 (16%) included *fellows*
- 17 (45%) included fully qualified *physicians/surgeons*
### Participant-Centred Methods

<table>
<thead>
<tr>
<th>Method (n)</th>
<th>Simulation</th>
<th>Clinical</th>
<th>Global-Rating</th>
<th>Task-Specific</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrasting-Groups (17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Borderline-Group (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1(1)*</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Generalized Examinee-Centred (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Receiver-Operator Curve (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4(1)*</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18</strong></td>
<td><strong>6</strong></td>
<td><strong>7</strong></td>
<td><strong>9</strong></td>
<td><strong>8</strong></td>
</tr>
</tbody>
</table>

*Values in parentheses indicate the number of studies using the method as a secondary means of standard setting.
## Item-Centred Methods

<table>
<thead>
<tr>
<th>Method (n)</th>
<th>Simulation</th>
<th>Clinical</th>
<th>Type of Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Global-Rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Task-Specific</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Both</td>
</tr>
<tr>
<td>Angoff (13)</td>
<td>12(1)*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>†Hofstee (6)</td>
<td>[6]</td>
<td>[0]</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>†Ebel (1)</td>
<td>[1]</td>
<td>[0]</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12</strong></td>
<td><strong>1</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>10</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>2</strong></td>
</tr>
</tbody>
</table>

*Values in parentheses indicate the number of studies using the method as a secondary means of standard setting.
†Hofstee and Ebel methods used in conjunction (or for comparison, in square brackets) with the Angoff method
**Judges**

- Only 16/37 (43%) studies describe judges as *trained in standard-setting AND content experts*
- 6 studies provided *no description of judges at all*
## Quality of the Evidence

<table>
<thead>
<tr>
<th>Method</th>
<th>n</th>
<th>MERSQI Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant-Centred</td>
<td>24</td>
<td>13.91 (12.50-15.50)</td>
</tr>
<tr>
<td>Contrasting-Groups</td>
<td>17</td>
<td>13.90</td>
</tr>
<tr>
<td>Borderline-Group</td>
<td>2</td>
<td>14.25</td>
</tr>
<tr>
<td>Generalized Examinee-Centred</td>
<td>1</td>
<td>14.50</td>
</tr>
<tr>
<td>Receiver-Operator Curve</td>
<td>4</td>
<td>13.63</td>
</tr>
<tr>
<td>Item-Centred</td>
<td>13</td>
<td>13.17 (11.00-14.05)</td>
</tr>
<tr>
<td>Angoff</td>
<td>13</td>
<td>13.17</td>
</tr>
<tr>
<td>Hofstee</td>
<td>6</td>
<td>13.41</td>
</tr>
<tr>
<td>Ebel</td>
<td>1</td>
<td>14.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>37</td>
<td><strong>13.67 (11.00-15.5)</strong></td>
</tr>
</tbody>
</table>

A score of ≥14/18 is considered ‘high quality’
## Procedures Assessed

<table>
<thead>
<tr>
<th>Setting</th>
<th>Medicine</th>
<th>Surgery</th>
<th>OBGYN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>MSK Ultrasound</td>
<td>Saphenofemoral Disconnection</td>
<td>Obstetric Ultrasonography</td>
</tr>
<tr>
<td></td>
<td>Colonoscopy</td>
<td></td>
<td>Vaginal Hysterectomy</td>
</tr>
<tr>
<td></td>
<td>Bronchoscopy*</td>
<td></td>
<td>Laparoscopic Hysterectomy</td>
</tr>
<tr>
<td>Simulation</td>
<td>Thoracocentesis</td>
<td>Basic Surgical Skills</td>
<td>Vaginal Surgery</td>
</tr>
<tr>
<td></td>
<td>Paracentesis</td>
<td>Cataract Surgery</td>
<td>Robotic Surgery</td>
</tr>
<tr>
<td></td>
<td>Endobronchial Ultrasound</td>
<td>Hip Fracture</td>
<td>Laparoscopic Surgery</td>
</tr>
<tr>
<td></td>
<td>Bronchoscopy*</td>
<td>Knee Arthroscopy</td>
<td>Colonoscopy</td>
</tr>
<tr>
<td></td>
<td>Bedside Procedures</td>
<td>Robotic Surgery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vascular Line Insertion</td>
<td>Colonooscopy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lumbar Puncture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colonoscopy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Study used both simulation and clinically obtained video footage in assessment
†One study used both surgical and obstetrics and gynaecology participants
‡Two studies assessing colonoscopy skill used surgeon-participants. MERSQI, Medical Education Research Quality Index.
Summary

• Studies carried out predominantly in simulation setting
  ➢ Allows for standardization of task
• Contrasting groups most commonly used method
• Inconsistent description of judges
• Quality of studies limited by:
  ➢ Validity evidence lacking
Moving Forward

Relative standards = Formative assessments
Absolute standards = Summative assessments

Incorporation of technical skill assessments into high-stakes examinations

Limit subjectivity in credentialing process

Ability predicts patient outcomes

Absolute standards are feasible in procedural assessment
Setting Performance Standards for Technical and Nontechnical Competence in General Surgery

Peter Szasz, MD,* Esther M. Bonrath, MD, PhD,*† Marisa Louridas, MD,* Andras B. Fecso, MD,*
Brett Howe, MD, MEd, FRCSC,‡ Adam Fehr, MD,§ Michael Ott, MD, MSc, FRCSC, FACS,‡
Lloyd A. Mack, MD, FRCSC, FACS,§ Kenneth A. Harris, MD, FRCSC, FACS,¶ and
Teodor P. Grantcharov, MD, PhD, FACS*

PGY1-5 performing laparoscopic cholecystectomy across 3 hospitals
Technical (OSATS) and non-technical (OSANTS) standards set
Judges fully trained and calibrated with anchor points
Our Current Work

• Setting absolute standards in surgeon technical performance in robotic surgery
  > Analysis of prospectively collected intraoperative robotic prostatectomy surgical video
    » Most commonly performed robotic surgical procedure
  > Contrasting Groups Method
    » Using metrics of global technical rating scales and assessment of surgeon error
    » Utilizes content expert judges
  > Linking performance metrics to functional/oncological outcomes
    » What role to surgeon factors play?
    » Can we set standards that ensure/enhance patient safety?
References


Help us improve. Your input matters.

• Download the ICRE App,
• Visit the evaluation area in the Main Lobby, near Registration, or
• Go to: http://www.royalcollege.ca/icre-evaluations to complete the session evaluation.

Aidez-nous à nous améliorer. Votre opinion compte!

• Téléchargez l’application de la CIFR
• Visitez la zone d’évaluation dans le hall principal, près du comptoir d’inscription, ou
• Visitez le http://www.collegeroyal.ca/evaluations-cifr afin de remplir une évaluation de la séance.

You could be entered to win 1 of 3 $100 gift cards.
Vous courrez la chance de gagner l’un des trois chèques-cadeaux d’une valeur de 100.
Additional Slides
## Item-Centred Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Open Surgery*</th>
<th>Endoscopic/arthroscopic</th>
<th>Laparoscopic/Robotic</th>
<th>Bedside Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant-Centred</td>
<td>6</td>
<td>9</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Contrasting-Groups</td>
<td>4</td>
<td>9</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Borderline-Group</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Generalized Examinee-Centred</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Receiver Operator Curve</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

*Two studies assessed a combination of open and laparoscopic surgical skills, and are included here*
## Item-Centred Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Open Surgery*</th>
<th>Endoscopic/arthroscopic</th>
<th>Laparoscopic/Robotic</th>
<th>Bedside Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item-Centred</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Angoff</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>†Hofstee</td>
<td>[0]</td>
<td>[0]</td>
<td>[0]</td>
<td>[6]</td>
</tr>
<tr>
<td>†Ebel</td>
<td>[0]</td>
<td>[0]</td>
<td>[0]</td>
<td>[1]</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6</strong></td>
<td><strong>9</strong></td>
<td><strong>10</strong></td>
<td><strong>12</strong></td>
</tr>
</tbody>
</table>

*Two studies assessed a combination of open and laparoscopic surgical skills, and are included here
†Hofstee and Ebel methods used in conjunction (or for comparison, in square brackets) with the Angoff method