The Inter-Rater Reliability of Technical Skills Assessment and Retention of Rater Training

Nada Gawad, Amanda Fowler, Richard Mimeault, Isabelle Raiche

Presenting author: Nada Gawad
Resident Physician, Division of General Surgery
Surgeon Scientist/Clinician Investigator Program
Department of Innovation in Medical Education
Introduction

• Valid and reliable assessment of trainees

• Tool development > reliability in practice

• Minimal rater training

• Global Operative Assessment of Laparoscopic Skills (GOALS)

• Optimal rater training unknown
Research Questions

1. What is the inter-rater reliability between surgeons with different practice backgrounds and experience?

2. What extent of rater training is necessary to achieve good inter-rater reliability between surgeons with different experience using GOALS?

3. Is rater-training retained over periods of non-use?
Methods

Procedure & Participants

• 33 simulated lap chole videos

• Surgeon assessors
  – Part 1: 5 surgeons, 3 videos
  – Part 2: 2 surgeons, 30 videos
Methods
Global Operative Assessment of Laparoscopic Skills (GOALS)

- Depth Perception
- Bimanual Dexterity
- Efficiency
- Tissue Handling
- Autonomy

Video-rating skills vs. technical skills
Methods

Study Design: Part 1

- Five-minute training video
- Independent video scoring
Methods

Study Design

Rater training: video + in-person discussion

In-depth rater discussion

1.5 months  1 month  1.5 months  1 month  2 months

Scoring #1  Scoring #2  Scoring #3  Scoring #4  Scoring #5

Figure 1. Timeline depiction of rater training and scoring sessions.
Methods

Analysis

- Videos
  - Unedited
- Intraclass correlations (ICC) and 95% CI
  - Mean-rating (k=2), absolute agreement, two-way random effects model
  - Reference:
    - Poor: <0.4
    - Moderate: 0.4-0.6
    - Good: 0.6-0.8
    - Excellent >0.8
Results

Part 1

- Mean score 12.4 (95% CI -2.31-0.98)
- ICC 0.26
### Results

**Part 2**

Table 1. Interrater reliabilities between the two raters for each scoring session

<table>
<thead>
<tr>
<th>Scoring Session</th>
<th>Number of videos</th>
<th>Mean Score</th>
<th>ICC</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>11.8</td>
<td>0.76</td>
<td>-0.35 to 0.99</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>11.2</td>
<td>-0.17</td>
<td>-2.70 to 0.85</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>9.6</td>
<td>0.74</td>
<td>-3.14 to 0.97</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>9.5</td>
<td>0.70</td>
<td>0.91 to 4.40</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>12.3</td>
<td>0.04</td>
<td>-0.88 to 0.75</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Discussion

- Multiple effective ways of training raters
- Not all training is equal
- Re-training required

- Assessment tools are ineffective without rater training
- Surgeon variation

- Limitations:
  - Two surgeons
  - Improved familiarity over time
Conclusions

• Good inter-rater reliability can be achieved with various methods of rater training
• Re-training is needed after periods of non-use to ensure reliable assessment.

Thank you!

Questions?