Background and Objectives

- Family Medicine trainees are increasingly taught in competency-based programs.
- Competence of clinical performance is frequently assessed using standardized patients in Objective Structured Clinical Examinations (OSCEs) for both formative and summative purposes.
- For both educational and legal reasons assessments by OSCEs must show they are valid, reliable and have transparent standard-setting processes.
- This systematic review surveys the quality of OSCEs as reported in the literature.

Search

- MEDLINE, PsycINFO, ERIC, ERC, Research & Development Resource Base (University of Toronto) from inception to June 2014
- Updated search in MEDLINE and PubMed single citation matcher to October, 2014
- 2 authors independently assessed all titles, abstracts and full texts, and abstracted data
- 3362 abstracts/titles
- 207 full-text articles assessed for eligibility
- 18 studies included

Method

- Articles were studied and information was extracted regarding population, intervention, comparison, outcome measures, and study design
- Evidence of validity and reliability studies was documented
- Standard-setting was studied in summative assessments

Results

Validity in 18 Studies 1987-2009

<table>
<thead>
<tr>
<th>Types of Validity</th>
<th>Number of Studies Which Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>14</td>
</tr>
<tr>
<td>Construct</td>
<td>8</td>
</tr>
<tr>
<td>Criterion / Concurrent / Predictive</td>
<td>5</td>
</tr>
</tbody>
</table>

Reliability

<table>
<thead>
<tr>
<th>Types of Reliability</th>
<th>Number of Studies Which Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>11</td>
</tr>
<tr>
<td>Inter-rater</td>
<td>6</td>
</tr>
<tr>
<td>Interstation</td>
<td>2</td>
</tr>
<tr>
<td>Inter-Item Correlations</td>
<td>2</td>
</tr>
<tr>
<td>Generalizability</td>
<td>4</td>
</tr>
</tbody>
</table>

Standard Setting

- 6 used an arbitrary cut-off point* or an empirical (normative, relative) method**
- 1 compared an empirical (normative) method to a rational (criterion-based) method***

* For example 60% (Garcia 2002)
** For example 2 standard deviations below mean
*** absolute method based either on characteristics of examination questions or on characteristics of the examinees

Examples of OSCEs and Results

1. Brailovsky et al. 1992
- 235 family medicine residents in four University centres in Québec.
- 24 clinical cases in a 40 station OSCE which was part of the 1991 certification examination
- Pass level set at 2 standard deviations below the mean-passing rate 95%

Examples of OSCEs and Results Continued

2. Friedlich et al. 2001
- 33 family practice residents randomly selected from 176 University of Toronto family medicine residents
- 8 station OSCE to assess performance of minor surgical office procedures
- Results compared to 14 senior surgical residents
- Average scores: PGY-1 family medicine residents 60%; PGY2 family medicine 64%; surgery residents 87%

3. Hamadeh 1993
- 31 Family Medicine residents at 3 different levels of training
- 10 station OSCE
- Compared: scores of first, second and third year residents; and OSCE results to ABFP ITERs
- Increasing scores with training year: Year 1: 6.93, year 2: 7.49, year 3: 8.25
- Significant correlation with ABFP ITERs and residents' monthly rotation evaluations

4. Regehr et al 1999
- 33 clinical clerks and 15 2nd year Family Medicine residents
- 5 station OSCE (12 minute patient encounter, 5 minute written post-encounter probe)
- Global rating scales compared to checklists generated by standardized patients and by MD examiners; performance scores of clerks compared to residents
- Residents scores were higher than clerks for both MD and standardized patient examiners
- No statistically significant differences between global rating and checklist scores
- Resident checklist scores: by physician examiner 67.99, by standardized patients 70.15
- Resident global scores: by physician examiner 82.93; by standardized patients 76.35
- Significant difference between MD and SP global scoring but not checklist scores

Conclusions

- There are a limited number of studies with robust validity and reliability assessments of family physician trainees' performance on Objective Structured Clinical Examinations.
- Content validity and internal reliability were most commonly measured, but not universally and generalizability analyses were seldom done.
- Summative assessments used a normative standard or an arbitrary cut-off for convenience, both of which are questionable for competency-based programs in the future.
- Criterion-based standard-setting methods are more appropriate in competency-based curricula when assessing performance.
- Standards should be based on guidelines and systematic reviews demonstrated to improve patient outcomes.
- Because of the impact of assessments on patients, residents and faculty, all OSCEs should have fully documented validity and reliability.