How can Technology Improve the Assessment of Medical Students, Residents, and Practicing Physicians?

John (Jack) R. Boulet, PhD FSSH
Vice President, Research and Data Resources
ECFMG/ FAIMER

AAHCI MENA Regional Conference, Beirut, Lebanon
Transformation of Medical Education in the New Era
September, 2018
Declarations

• I have no conflicts of interest

• I am not a physician, but have worked in a medical school

• Information on automated item generation is courtesy of the Medical Council of Canada (MCC) where I serve as co-chair of their Research Advisory Committee

• I am part of the Clinical Skills Evaluation Collaboration (CSEC)
New Developments (Technology)
Outline of Talk

• Technology in education and assessment
  – Evolving competencies
• Test administration
  – Adaptive testing, automated item generation
• Scoring (AI applications)
  – Policy capturing, NLP
• Assessing practicing physicians
• Looking to the future
Education and Assessment

• Evolution of learning models & technologies not completely mirrored by similar changes in educational assessment
  – Assessment is still mostly “physician-centric, individually tailored” with traditional modalities (MCQs, OSCEs, static/linear representations of materials on screen)
  – Educational assessment must evolve alongside learning models/technologies
The Practice of Medicine

• Competencies that will be critical to assess in medical education (AMA)
  – Inquiry and improvement (“dealing with uncertainty with big data”)
  – Interdependency (“working in teams”)
  – Information management ("informatics")
  – Interest and insight (“patient-centered care”)
  – Involvement (“adding life to years and not just years to life”)

• How well do we assess these competencies?
Test Administration and Scoring

• Both selected constructed response items can be administered on the computer
  – Secure, efficient, adaptive

• Technology-enhanced scoring will continue to evolve and grow
  – Reliable, efficient, “economical”
  – Automated feedback
Automated Item Generation (AIG)

- AIG is the process of using models to generate test items with the aid of computer technology.
- AIG uses a three-stage process for generating items where the cognitive mechanism required to solve the items is identified and manipulated using computer technology to create new items.
Automated Item Generation (AIG)
Automated Item Generation: Cognitive Map

A structure of problem solving knowledge for issues related to post-operation fevers.

- **Problem and Scenarios**
  - Urinary Tract Infection (UTI)
  - Asystole
  - Wound Infection (W)
  - Pneumonia (P)
  - Deep Vein Thrombosis (DVT)
  - Deep Space Infection (DSI)

- **Sources of Information**
  - Timing of Fever
  - Physical Examination
  - Type of Surgery

- **Features**
  - 1-2 Days
    - Fever
      - Normal range of body temp.
      - No guarding or rebound
      - No Constraint
    - Abdominal Examination
      - Element
        - Normal
        - Tenderness
      - Constraint
      - DSG Present
  - 2-3 Days
    - Guarding and Rebound
      - Element
      - No guarding or rebound
      - Constraint
      - DSG Present
      - UUT: Very Rare
      - UTI: Very Likely
    - Fever
      - Normal range of body temp.
      - No Constraint
  - 4-5 Days
    - Element
      - No guarding or rebound
      - Constraint
      - UUT: Unlikely
      - Abdominal Examination
        - Element
          - Normal
          - Tenderness
        - Constraint
        - DSG: Unlikely

- **Types of Surgery**
  - Gastrectomy
    - Element
      - Right Hemicolectomy
      - Left Hemicolectomy
      - Appendectomy
    - Element
      - UUT: Very Likely
      - UUT: Very Likely
  - Laparoscopic Cholecystectomy
    - Element
      - Right Hemicolectomy
      - Left Hemicolectomy
      - Appendectomy
Automated Item Generation: Cognitive Map

• Complex diagram that highlights knowledge, skills and content required to make a medical diagnosis or manage a patient.

• The model includes three key activities:
  1. Identifying **THE PROBLEM** (i.e., post-operative fever)
  2. Specifying **SOURCES OF INFORMATION** required to diagnose the problem (i.e., type of surgery)
  3. Describing **KEY FEATURES** within each information source (e.g., fever) needed to create different instances of the problem
Automated Item Generation: Item Model

• Next, we create item models using the cognitive map content; an item model is a template or a mould of the assessment task (i.e., it’s a target where we want to place the content for the item)

• A 54-year-old woman has a <TYPE OF FEVER>. On post-operative day <TIMING OF FEVER>, the patient has a temperature of 38.5°C. Physical examination reveals <PHYSICAL EXAMINATION>. Which one of the following is the best next step?

<TYPE OF FEVER> Gastrectomy; right hemicolectomy; left hemicolectomy; appendectomy; laparoscopic cholecystectomy
<TIMING OF FEVER> 1 to 6 days
PHYSICAL EXAMINATION> Red and tender wound; guarding and rebound; abdominal tenderness, calf tenderness
Automated Item Generation with IGOR

• After the item model is specified, the information is systematically combined to produce new items

• To accomplish this complex, combinatoric task, an item generation software called IGOR (Item GeneratOR) was created
Automated Item Generation: Lessons Learned

- **Usefulness of distractors:**
  - Early attempts at AIG generated items which on occasion had (medically) non-plausible distractors

- **Engine was recoded to allow greater control over distractor combinations and more complex relationships**
Automated Item Generation: Lessons Learned

• Complexity of coding:
• Code was complex and not amenable to on-the-fly revisions
• Period of several weeks required to recode the maps and regenerate items for review
• To resolve this problem, the Medical Council of Canada developed an interface (*Item Butler*) that allows test committee members to create their own cognitive models, revise them on-site and generate/revise samples of items for review
A 68 Kg, 50-year-old security guard was during an altercation with another man. He is immediately brought to the hospital where his vitals are as follows: blood pressure $\text{BP}$, heart rate, and respiratory rate. He has $\text{D: AIR_ENTRY}$ with $\text{C: HEART_SOUNDS}$. His neck veins are $\text{B: NECK VEINS}$ and his trachea $\text{A: TRACHEA}$.

What is the most likely diagnosis?

**Objective:** Chest Trauma Diagnosis

**Task:**

**STEP 2: CORRECT OPTIONS LIST**

- **A: TRACHEA**
  - A1: is midline
  - A2: is deviated to the right side

- **B: NECK VEINS**
  - B1: distended
  - B2: not distended
  - B3: not visible

- **C: HEART SOUNDS**
  - C1: faint heart sounds
  - C2: no heart sounds
  - C3: pronounced systolic ejection murmur
  - C4: hyperresonance on percussion

- **D: AIR ENTRY**
  - D1: good air entry
  - D2: decreased air entry to the bases
  - D3: good air entry bilaterally
  - D4: no air entry on the left side

- **E: BP**
  - E1: 70/60
  - E2: 85/78
  - E3: 88/80
  - E4: 95/70
  - E5: 90/50
Automated Item Generation: Next Steps

• Improving feedback to candidates
  – Systematizing elements of a cognitive map to provide “sliceable” multidimensional feedback according to blueprint dimensions, objectives, cognitive taxonomy, etc.

• Research using AIG with other item types
  – Assessing feasibility of using AIG to create cognitive maps for write-in & short-menu items (CDM cases)

• Modelling psychometric properties of AIG-based items
  – Using classification & regression tree analysis to help better model performance of AIG items in the future
Technology-Enhanced Scoring

- Can we use computers to score some or all of our assessment exercises?
  - Experts can be influenced by irrelevant variables when formulating their judgments
  - Scores may not adequately represent performance
  - Cost/logistics

Artificial Intelligence
Technology-Enhanced Scoring

• Explicit outcome measures (policy capturing)
  – Code actions and timing based on machine readable output

• Natural Language Processing (NLP)
  – Post simulation encounter exercises, essays

• Analysis of Talk Time
  – Speech recognition

• Automated Computer Vision Analysis
  – Empathy, rapport
Automated Scoring

- A software package infers automated scoring rules for specific problems based on examples of student work ratings/gradings provided by instructors.
- A number of algorithms can be used to create this automated scoring rule including
  - Linear/logistic regression, Decision Tree, SVM, Naive Bayes, etc.
- Supervised learning used extensively to grade essays, short-answer items as well as more complex computer-based patient management tasks.
Natural Language Processing

- Automated essay scoring is currently available ... and being used
- Can be applied to various assessments used in medical education
Value of Automated Scoring

• The appeal of AI assessment lies in its efficiency and consistency in applying the same criteria across students
• Possibility of offering immediate and detailed feedback on performance to students
Policy Capturing

- USMLE Step 3
  - Computer-based case simulations
- Mannequin-based simulations with machine readable transaction lists
  - Code action and timing based on computer output
- Other types of items?
  - Clinical Decision Making (CDM)
Automated Speech Recognition

- Analysis of talk time
  - Communication
    - Linguistic features, talk time ratios, etc.
- Coding of content
  - History taking
Automated Computer Vision Analysis

• Communication measures
  – Empathy, rapport

• Motion analysis
  – Gymnastics, diving
  – Physical examination or procedures
    • Demands specific performance scoring criteria
Assessment of Practicing Physicians

• For practicing clinicians, one can look at patient outcomes as a metric for quality of care
  – Continuous assessment and feedback
  – Maintenance of certification

• Demands comprehensive/structured data systems
  – Comparable EHRs
Discussion

• The physician’s role will be even more critical in regard to the introduction, evaluation and best use of these technologies in their role as “health advisor and knowledge navigator”

• From an assessment perspective, computer technology is not the problem but the solution

• The use of diagnostic, management, data mining and summarization algorithms is/will drastically alter(ing) medical education and assessment

• AI in medicine (and assessment) will be a “team sport” predicated on a set of new competencies (statistics, computer sciences, etc.)
Conclusion

• As the practice of medicine changes, so too must the assessment systems
  – Assessment drives education ... hopefully in the right direction

• Artificial intelligence offers the opportunity to make our assessments more relevant, valid, reliable and (perhaps) less costly