Council on Licensure, Enforcement and Regulation
2011 Annual Educational Conference

Pittsburgh Pennsylvania

Automated Scoring of Performance Tasks

Presenters:
F. Jay Breyer, ETS
Richard DeVore, AICPA
Ronald Nungester, NBME
Chaitanya Ramineni, ETS
Dongyang Li, Prometric

Promoting Regulatory Excellence

F Jay Breyer, PhD
Educational Testing Service

WHAT IS AUTOMATED SCORING?

Why Constructed Response Items?

- Constructed Response – examinee generates a response rather than selecting from presented options
- Challenges
  - Development and administration
  - Human scoring: recruitment, training, score quality, multiple raters
  - Score turnaround
  - Information/reliability relative to multiple-choice per unit time
- Demand
  - Construct coverage – address something that is valued and thought to be inadequately covered by MC
  - Face validity – real-world fidelity to naturalistic tasks is valued

3

What do we mean by automated scoring?

- Estimate an examinee's proficiency on the basis of "performance tasks" (writing, speaking, drawing, decision making, etc.), without direct human intervention
- Typically, the computer will be trained to identify features of task responses which are strongly predictive of human ratings, and will be optimized to maximize its agreement with human ratings

4

Why Automated Scoring?

- Time
- Cost
- Scheduling
- Consistency
- Performance Feedback
- Construct Expansion

Challenges of Automated Scoring

- Time for development
- Cost of development
- Consistency
- Lack of credentials (a résumé)
- Expectations of score users and public

5

What Can be Scored Automatically?

- Essays for Writing proficiency
- Short Text Responses
  - for Correct answers (concepts)
- Mathematics Tasks
  - Equations, Graph data responses, Quantitative values
- Spoken Language –
- Simulations

6

F. Jay Breyer, PhD
Educational *Testing Service*

A FRAMEWORK FOR EVALUATION AND USE OF AUTOMATED SCORING

7

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Our Framework

I. Consideration of Validity & Reliability Issues
   - Guided by theory
II. Empirical Evidence Supportive of Use
   - Held accountable
III. Policies for Implementation & Use
   - There is a need for guidelines and limits

8

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

I. Validity & Reliability Issues

- Validity:
  - Construct Relevance vs. Irrelevance
    - How well do extracted features fit with claims/important inferences?
    - Are there features extracted from the automated scoring engine that are proxies for the intended inferences?
      - More or less valued features act as proxies for the direct construct
  - Construct Representation vs. Underrepresentation
    - Are the features extracted by the automated scoring system sufficient to cover the important aspects of the performance for the intended claims?
      - Are there enough of them?
    - Are the extracted features too narrow?
      - e.g., Simply counting words

9

I. Validity &Reliability Issues

- Reliabilty:
  - Accuracy
    - How well do the automated scores agree with some analogous true-score substitute measure?
  - Consistency
    - Are automated scores consistent across tasks, raters, occasions?
  - An Example

10

II. Empirical Evidence to Support Use

- For Validity:
  - Gather evidence:
    - Are the features relevant to the claims?
      - (construct relevance vs. irrelevance)
    - Are the features too narrow or too broad?
      - (construct representation vs. underrepresentation)
    - Validity Studies
      - Factor Analytic studies, Multitrait-Multimethod, etc.

11

Empirical Evidence

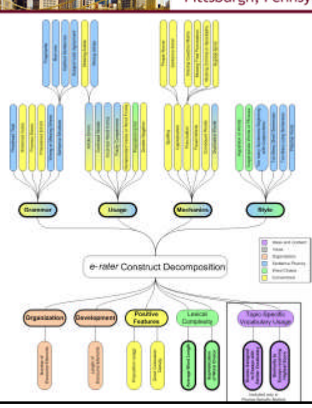
For writing:

•Do the features appear to capture what is important for scoring essays in this case?

Judgmental Process:

• The different colors map to different traits in the model

• The features are proxies for what is important in the construct.



II. Reliability & Validity

- For Reliability
  - Internal evidence
    - Agreement with some true-score substitute
      - We use human scorers
        - We look at agreement above chance
          - *Quadratic-weighted kappa*
    - Consistency
      - We use human scorers
        - *Correlation* of H & AS

13

II. Reliability & Validity

- For Reliability
  - Internal evidence
    - Degradation
      - Loss of accuracy or consistency when using automated scores compared to human scores
        - We look at (H1,H2)-(H,AS) for weighted *kappa* and *correlations*
    - Standardized Mean Difference

$$\sigma_{M\_difference} = \frac{\left[\overline{X}_{AS} - \overline{X}_{H}\right]}{\sqrt{\frac{SD^2_{AS} + SD^2_{H}}{2}}}$$

14

Some Caveats

- Use of weighted kappa, correlation, and human-human agreement are informative

	human1		human2			human1-automated			
	mean	sd	mean	sd	wtd k	mean	sd	std diff	wtd k
Average	3.85	0.96	3.86	0.96	0.74	3.85	0.95	0.00	0.76

- … but can be incomplete

	human1		human2			human1-automated			
	mean	sd	mean	sd	wtd k	mean	sd	std diff	wtd k
Subgroup	3.29	0.77	3.29	0.77	0.39	3.74	0.70	0.60	0.39

15

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

III. Policies

- When do humans intervene?
  - Advisories
    - When we *cannot* score a performance with automated scoring techniques
    - When we are suspicious automated score use is inappropriate
  - Threshold for adjudication
    - How much of a difference do you need to see before you require a human to take a look?
    - Thresholds vary in practice

16

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Examination Stakes

- Low Stakes
  - Practice environment
  - Learning environment
    - Used without human intervention
- Medium Stakes
  - Formative assessments where more than one measure is used
    - Used without human intervention with a subsample scored by humans for evaluation purposes
- High Stakes
  - Make or break examinations
    - Used as a contributing score along with human scores
    - Exceeding adjudication thresholds requires a second human score

17

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Finally

- Remember
  - We want to be
    - guided by theory
    - supported by evidence
      - It's not just agreement or correlation
      - Use appropriate evaluation metrics
      - Disaggregate tasks and subgroups
    - true to our policies
      - No one scoring solution will fit everything
      - Qualify which humans, under what circumstances and for which data

18

CLEAR 2011
Annual Conference

September 8-10
Pittsburgh, Pennsylvania

Ronald J. Nungester, PhD, Brian E. Clauser, EdD. Polina Harik, PhD
National Board of Medical Examiners

AUTOMATED SCORING OF SIMULATIONS IN MEDICAL LICENSURE

19

Council on Licensure, Enforcement and Regulation
2011 Annual Educational Conference

Pittsburgh Pennsylvania

Automated Scoring of Simulations in Medical Licensure

Presenters:

Ronald J. Nungester, PhD
Brian Clauser, EdD
Polina Harik, PhD
National Board of Medical Examiners

Promoting Regulatory Excellence

CLEAR 2011
Annual Conference

September 8-10
Pittsburgh, Pennsylvania

NBME Products and Services

- USMLE™
- Services for medical schools and students
- Services for healthcare organizations
- Services for practicing doctors
- International collaboration
- Research &development

USMLE

- Introduced by the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB) in 1992
- Sole examination pathway for allopathic medical licensure in the US
- Administered in three Steps
  - **Step 1: understanding of biomedical science**
  - **Step 2 (CK & CS): readiness for supervised graduate training**
  - **Step 3: readiness for unsupervised practice**

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

USMLE Simulations

- MCQs
  - Vignette Based
  - Pictorials
  - Multimedia (sound, video, animations)
- Computer-Based Case Simulations (Primum®)
- Standardized Patients
- Automated scoring applications in CCS and SPs

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Clinical Skills Examination

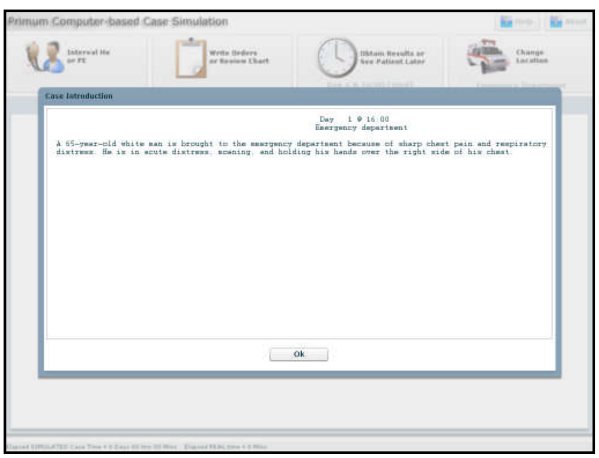
- Component of Step 2
- Prerequisite for Step 3
- 12 standardized patients
- 3 hurdles: English-language, communication, integrated care including Patient Notes
- 5 test sites - Houston, Chicago, LA, Atlanta, Philadelphia

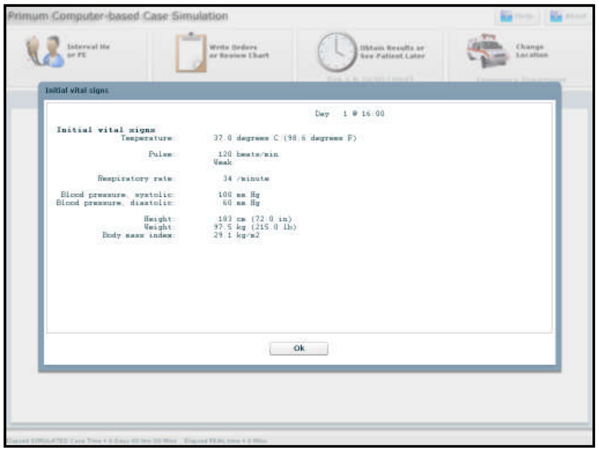
Clinical Skills Examination

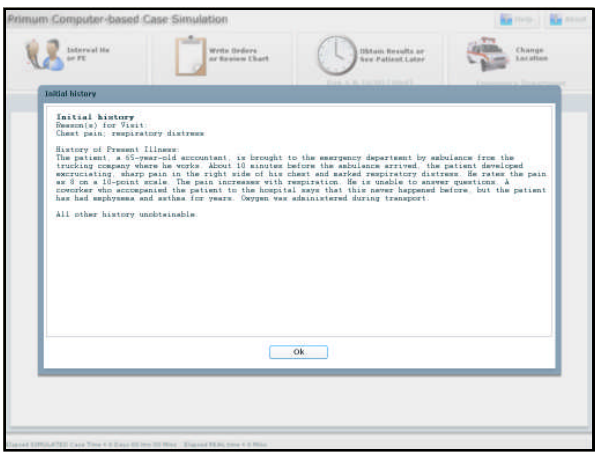
- Investigating automated scoring of PN
- Application of Natural Language Processing (NLP)
- Augment or replace physician raters
- Rule-based and regression-based scoring procedures being considered

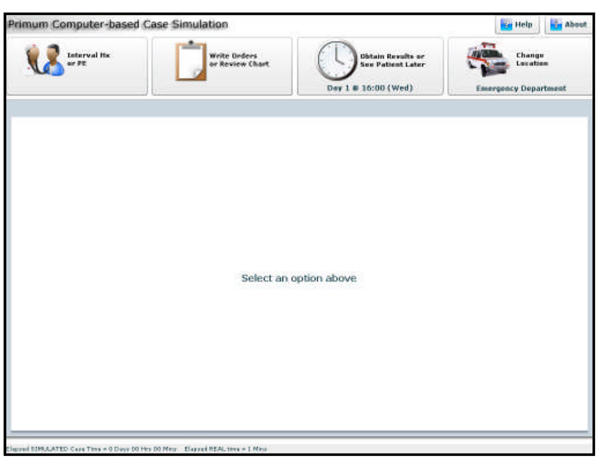
Primum® Clinical Case Simulations

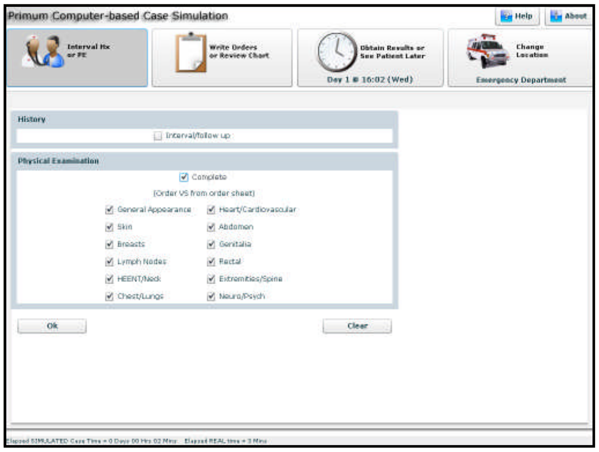
- Simulated environment allows observation of clinical management
- Observed behavior scored
- Dynamic
- Unprompted
- Free response
- Used in Step 3

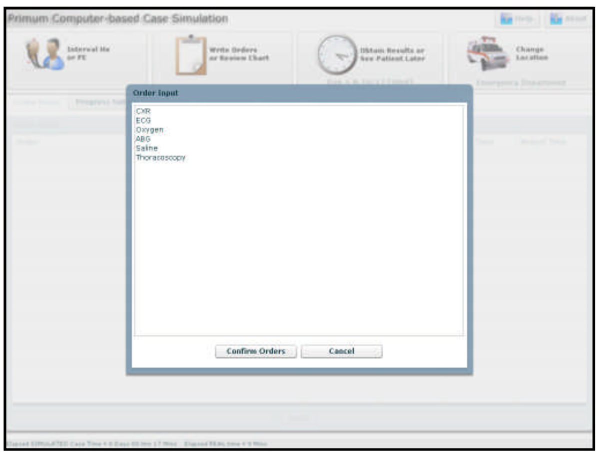


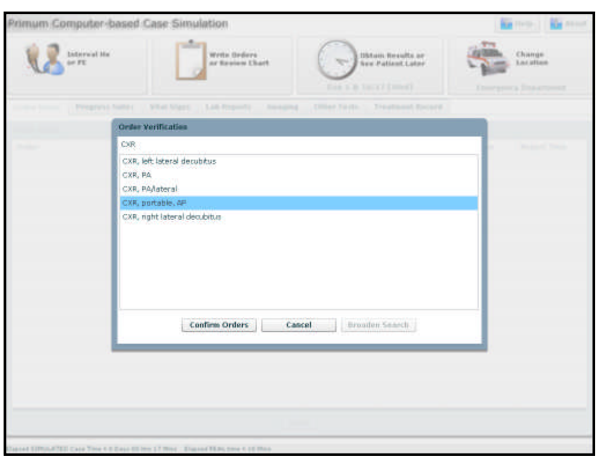

Primum Computer-based Case Simulation
Initial vital signs
Day 1 @ 16:00
Initial vital signs
Temperature: 37.0 degrees C (98.6 degrees F)
Pulse: 120 beats/min
Weak
Respiratory rate: 34 /minute
Blood pressure, systolic: 100 mm Hg
Blood pressure, diastolic: 60 mm Hg
Height: 183 cm (72.0 in)
Weight: 97.5 kg (215.0 lb)
Body mass index: 29.1 kg/m2
Ok

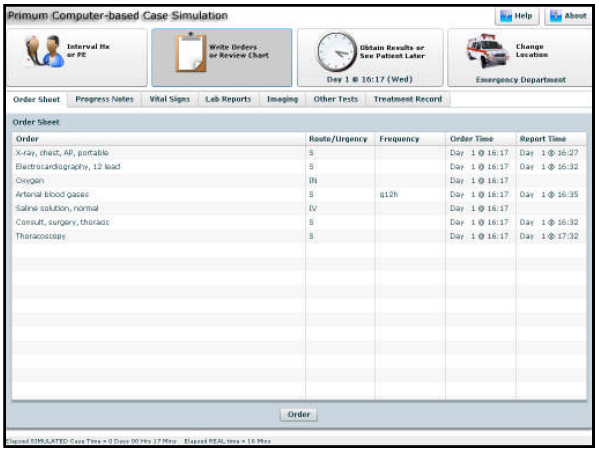

Primum Computer-based Case Simulation
Initial history
Initial history
Reason(s) for Visit:
Chest pain; respiratory distress
History of Present Illness:
The patient, a 65-year-old accountant, is brought to the emergency department by ambulance from the trucking company where he works. About 10 minutes before the ambulance arrived, the patient developed excruciating, sharp pain in the right side of his chest and marked respiratory distress. He rates the pain as 8 on a 10-point scale. The pain increases with respiration. He is unable to answer questions. A coworker who accompanied the patient to the hospital says that this never happened before, but the patient has had emphysema and asthma for years. Oxygen was administered during transport.
All other history unobtainable.
Ok

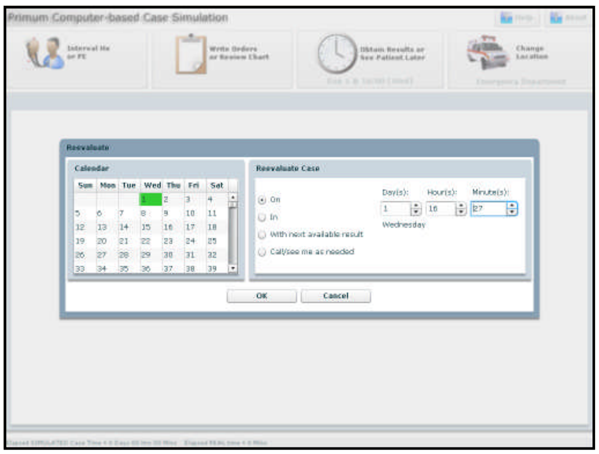

Primum Computer-based Case Simulation
Help
About
Interval Hx or PE
Write Orders or Review Chart
Obtain Results or See Patient Later
Day 1 @ 16:00 (Wed)
Change Location
Emergency Department
Select an option above

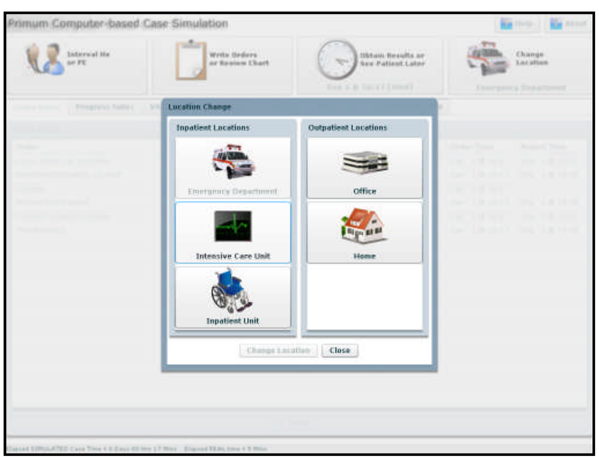

Primum Computer-based Case Simulation
Help
About
Interval Hx or PE
Write Orders or Review Chart
Obtain Results or See Patient Later
Day 1 @ 16:02 (Wed)
Change Location
Emergency Department
History
Interval/follow up
Physical Examination
Complete
(Order VS from order sheet)
General Appearance
Heart/Cardiovascular
Skin
Abdomen
Breasts
Genitalia
Lymph Nodes
Rectal
HEENT/Neck
Extremities/Spine
Chest/Lungs
Neuro/Psych
Ok
Clear

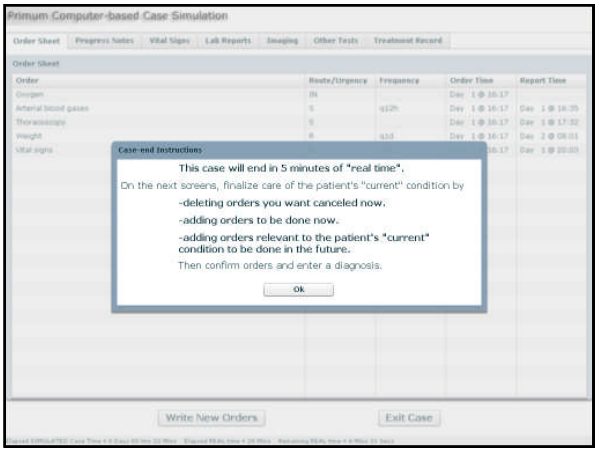

Primum Computer-based Case Simulation
Order Input
CXR
ECG
Oxygen
ABG
Saline
Thoracoscopy
Confirm Orders
Cancel

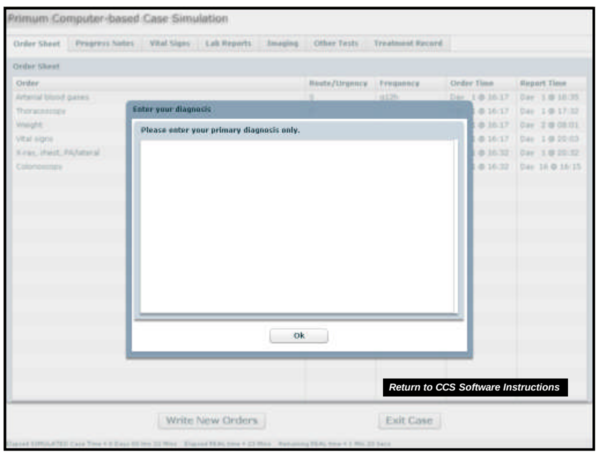

Primum Computer-based Case Simulation
Order Verification
CXR
CXR, left lateral decubitus
CXR, PA
CXR, PA/lateral
CXR, portable, AP
CXR, right lateral decubitus
Confirm Orders
Cancel
Broaden Search


Primum Computer-based Case Simulation
Help
About
Interval Hx or PE
Write Orders or Review Chart
Obtain Results or See Patient Later
Day 1 @ 16:17 (Wed)
Change Location
Emergency Department
Order Sheet
Progress Notes
Vital Signs
Lab Reports
Imaging
Other Tests
Treatment Record
Order Sheet
Order
Route/Urgency
Frequency
Order Time
Report Time
X-ray, chest, AP, portable S Day 1 @ 16:17 Day 1 @ 16:27
Electrocardiography, 12 lead S Day 1 @ 16:17 Day 1 @ 16:32
Oxygen IN Day 1 @ 16:17
Arterial blood gases S q12h Day 1 @ 16:17 Day 1 @ 16:35
Saline solution, normal IV Day 1 @ 16:17
Consult, surgery, thoracic S Day 1 @ 16:17 Day 1 @ 16:32
Thoracoscopy S Day 1 @ 16:17 Day 1 @ 17:32
Order
Elapsed SIMULATED Case Time = 0 Days 00 Hrs 17 Mins Elapsed REAL time = 16 Mins


Primum Computer-based Case Simulation
Reevaluate
Calendar
Sun Mon Tue Wed Thu Fri Sat
Reevaluate Case
On
In
With next available result
Call/see me as needed
Day(s):
Hour(s):
Minute(s):
Wednesday
OK
Cancel


Primum Computer-based Case Simulation
Location Change
Inpatient Locations
Outpatient Locations
Emergency Department
Office
Intensive Care Unit
Home
Inpatient Unit
Change Location
Close





CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

Sample Transaction List

Ordered	Action	Seen
1@16:00	HEENT/neck	1@16:11
1@16:00	Cardiac examination	1@16:11
1@16:00	Chest/lung examination	1@16:11
1@16:11	X-ray, portable	1@16:31
1@16:11	Arterial blood gases	1@16:26
1@16:11	Electrocardiography, 12 lead	1@16:41
1@16:11	Oxygen by mask	
1@16:14	Patient Update ("More difficulty breathing")	
1@16:14	Needle thoracostomy	1@16:19
1@16:24	Chest tube	
1@16:30	Patient Update ("Patient feeling better")	
1@16:30	Chest/lung examination	1@16:31

Action Categories

- Beneficial Actions
  - Least important
  - More important
  - Most important
- Detractors
  - Non-harmful
  - Risky
  - Extremely Dangerous
- Timing/Sequence

Initial Scoring Approaches

- Raw Score (Unit Weighting)
- Rule-based policy capturing
- Regression-based policy capturing

Rule-Based Policy Capturing

- Experts articulate rules for required levels of performance for each score category
- Rules operationalized by identifying the specific combinations of actions required for each score level

Example: Rule-based Scoring

- Logical statements mapping patterns of performance into scores
- Reflected case-specific scoring key
- Example
  - Dx + Rx +Mn, no non-indicated actions = 9
  - Dx + Rx, no non-indicated actions = 7
  - Dx, no Rx = 2

Regression-Based Scoring

- Experts review and rate a sample of transaction lists
- Regression equation produced for each case
  - Dependent measure
    Mean expert rating
  - Independent measures
    Count of items within each action category
- Algorithms produce scores that approximate the ratings that would have been produced by content experts

Estimated Regression Weights

Variable	Weight
Beneficial - Most	1.50
Beneficial - More	0.75
Beneficial - Least	0.20
Non-harmful	-0.05
Risky	-1.10
Extremely Dangerous	-2.00
Timing	1.30

Weighted score=$1.5*B_{most}+\ldots-2*ED+1.3*TM$

Correlations between Ratings and Scores

Case	Raw Score	Regression-based Score	Rule-based Score
1	.76	.81	.77
2	.66	.91	.85
3	.78	.89	.87
4	.80	.88	.84
5	.77	.84	.69
6	.71	.86	.87
7	.54	.79	.79
8	.78	.95	.86

CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

Scoring Approaches

- Rule-Based Scores
- Regression-Based Weights
- Unit Weights
- Fixed Weights
- Averaged Weights

CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

Scoring Weights

- Unit Weights
  Score = Most Important + Less Important + Least Important - Inappropriate - Risky - Harmful
- Fixed Weights
  Score = 3*Most Important + 2*Less Important + Least Important - Inappropriate - 2*Risky - 3*Harmful
- Averaged Weights
  Score = W1*Most Important + W2*Less Important + W3*Least Important - W4*Inappropriate - 5*Risky- W6*Harmful

Score-Rating Correlations

averaged across 18 cases

	Regression-based	Rule-based	Unit weights	Fixed weights	Average weights
Mean	0.86	0.85	0.75	0.75	0.75
Median	0.87	0.86	0.75	0.76	0.79
SD	0.05	0.08	0.06	0.08	0.13

Score Reliability

	Regression-based	Rule-based	Unit weights	Fixed weights	Average weights
form1	0.39	0.27	0.47	0.46	0.45
form2	0.46	0.42	0.49	0.49	0.47
form3	0.42	0.36	0.47	0.45	0.48
Mean	0.42	0.35	0.48	0.47	0.47

Correlations with Multiple Choice Score

	Regression-based	Rule-based	Unit weights	Fixed weights	Average weights
Observed Correlations					
form1	0.31	0.30	0.34	0.34	0.26
form2	0.39	0.42	0.41	0.40	0.35
form3	0.34	0.32	0.37	0.33	0.18
Mean	0.35	0.35	0.37	0.36	0.27
Corrected Correlations					
form1	0.51	0.61	0.51	0.51	0.41
form2	0.60	0.68	0.61	0.60	0.53
form3	0.55	0.55	0.56	0.52	0.27
Mean	0.55	0.61	0.56	0.54	0.40

Automated Scoring

- Provides a good approximation of expert ratings
- Regression-based scoring does not require experts to be explicit about their rating policies
- Rule-based scoring allows for explicit evaluation of the scoring process
- Rule-based scoring may be more efficient than regression based procedures

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Automated Scoring

- Identifying and quantifying components of performance is more important than weighting them in creating a score
- Case-specific scoring models better approximate ratings than do generic models
- Rule-based scoring may be more preferable for practical and theoretical reasons
- Higher apparent reliability may result from measuring construct-irrelevant or secondary traits
- Gradual improvements in case and key development warrant re-examination of scoring procedures over time

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Automated Scoring

- As reliable as scores produced by expert raters
- Developing the scoring algorithms for regression-based scoring may be resource intensive
- Regression procedures may not adequately model unusual response patterns

Automated Scoring

- Highly efficient
  - More than 2,500,000 cases have been scored electronically
  - Expert review and scoring of this same number of performances would have required more than 100,000 hours of rater time

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

Automated Scoring of Simulations in Medical Licensure

- Ronald J. Nungester, PhD
  - Senior Vice President, Professional Services
    - National Board of Medical Examiners
    - 3750 Market Street
    - Philadelphia, PA 19104
    - rnungester@nbme.org
- For additional information and sample cases:
    - www.nbme.org
    - www.usmle.org

CLEAR 2011
Annual Conference
September 8-10
Pittsburgh, Pennsylvania

SCORING SHORT TEXT RESPONSES FOR CONTENT IN A LICENSURE EXAMINATION

Richard DeVore, Ed.D.
Joshua Stopek, CPA
AICPA

57

The Uniform CPA Examination

- 60 percent MCQ testing the body of knowledge for CPAs
- 40 percent Task-based Simulations (TBS)
  - Designed to replicate on-the-job tasks of the entry-level CPA
  - Tasks comprise 6 to 8 measurement opportunities (MO)

Measurement Opportunities

- MOs utilize several task formats
  - Constructed-response, numerical entry (scored objectively)
  - "Mega-multiple choice" selection (scored objectively)
  - Combination of the former two
  - "Research" item type (scored objectively)
  - Constructed-response, writing sample (scored by e-Rater)

C-Rater Study

- This study undertaken to determine whether c-Rater can reliably and accurately score constructed response answers for content
  - This might allow replacement of some selection answer types
  - Would improve the face validity of the TBS and remove the guessing factor
  - Would remove barrier to scoring true constructed response without human involvement

C-Rater Study

- TBSs were chosen from ones used for the writing sample
- All intended answers were taken directly from authoritative literature
- Authoritative literature was not available
- Exercises were not speeded

61

C-Rater Study

- All prompts assessed several concepts
  - Four prompts expected several concepts in one answer
  - One prompt was broken into three separate concepts
  - All concepts were supported by the authoritative literature
  - Sample responses were generated by SMEs

62

The Population

- CPA-bound Students
- Five Universities

College year	Total
Graduate	57
Junior	22
Senior	173
Sophomore	1
Grand Total	253

63

Prompt1

- When determining whether to accept an audit engagement previously performed by another firm, what information should your firm request from the predecessor firm?
- C-Rater Concepts (1 point per concept)
  - C1: Information that might bear on the integrity of the management OR information bearing on the integrity of the management OR Information about the integrity of the management (Anything that shows the management is dishonest)
  - C2: Any disagreements/arguments/conflicts/issues/differences with management
  - C3: Communications regarding/about fraud by the client OR Communications regarding/about illegal acts by the client
  - C4: Communications about significant deficiencies (in internal control)
  - C5: Communications about material weaknesses (in internal control)
  - C6: The reason for/why the change in auditors

64

Results Item 1

- Statistics are Quadratic-Weighted Kappas that look at the agreement over chance
  - Like a correlation except the further apart the two rating, the more the statistic degrades
  - Criterion for use is 0.70
- Item #1 meets the Criterion
- Question 1 asked for specific information

Item 1:

Set	H1:H2	H1:C	H2:C
Development	0.86	0.84	0.86
X-Evaluation	0.89	0.87	0.76
Blind	0.91	0.77	0.79

65

Prompt 2

- Analytic procedures are employed in the three phases of an audit (the beginning of the audit, during the audit, and at the end of the audit) for three distinct purposes. In each of the boxes below, briefly describe the purpose of analytic procedures for the indicated phase of the audit.
- C-Rater Concepts
  - A. In the beginning of the audit:
    - C1: To assist in the planning of the nature, timing and extent of audit procedures
  - B. During the audit:
    - C2: As substantive tests of audit assertions
  - C. At the end of the audit:
    - C3: To evaluate the overall financial statement presentation

66

What if humans cannot agree?

- When humans cannot agree
  - It makes little sense to build item models
  - Each Item requires its own model

Item 2a

Set	H1:H2	H1:C	H2:C
Development	0.44	0.47	0.65
X-Evaluation	0.57	0.36	0.48
Blind	0.34	0.18	0.40

Item 2b

Set	H1:H2	H1:C	H2:C
Development	0.49		
X-Evaluation	0.64		
Blind	0.65		

Item 2c

Set	H1:H2	H1:C	H2:C
Development	0.30		
X-Evaluation	0.34		
Blind	0.28		

67

Analysis of Prompts 1 & 2

- Item 1 worked because the response required specific types of information
- Item 2 failed because the meanings of the expected concepts were somewhat ambiguous, and SMEs differed on the appropriateness of candidate responses
- Item 2 also involves some “contra concepts” that may have been missed by SMEs or c-Rater

68

Prompts 3, 4, & 5

- (3) During the planning phase of the audit of MixCorp, the audit manager asked for assistance in determining the procedures to perform over inventory. What documents should be examined to test the rights and obligations assertion of MixCorp’s inventory?
- (4) Willow Co. is preparing a statement of cash flows and needs to determine its holdings in cash and cash equivalents. List three examples of cash equivalents that Willow should remember to include.
- (5) Give two examples of circumstances under which long-lived assets should be assessed for potential impairment.

69

Items 3, 4 & 5

- Again the statistics are Quadratic-Weighted Kappas that look at the agreement over chance
- Item 3 is good both in terms of HH agreement and H & c-rater agreement
- Item 4 is good and actually learns from the xval data set improving over the development stage.
- Item 5 c-rater has challenges in scoring this item

Item 3	Set	H1:H2	H1:C	H2:C
	Development	0.77	0.80	0.75
	X-Evaluation	0.81	0.86	0.84
	Blind	0.75	0.84	0.77

Item 4	Set	H1:H2	H1:C	H2:C
	Development	0.83	0.51	0.58
	X-Evaluation	0.82	0.70	0.75
	Blind	0.78	0.71	0.72

Item 5	Set	H1:H2	H1:C	H2:C
	Development	0.77	0.50	0.59
	X-Evaluation	0.77	0.54	0.54
	Blind	0.57	0.49	0.55

70

Analysis of Prompts 3,4, & 5

- Items 3 & 4 worked because the responses required limited sets of quite specific examples
- Item 5 failed because the expected concepts were classes of items, but candidates responded with specific examples, each of which had to be interpreted and judged independently

71

Findings

- Response space has to be limited (Candidates can be verbose)
- Preparation of prompts required extensive refinement to make them amenable to c-Rater scoring
  - Prompts could not allow for judgment and related explanation of thought
  - Concepts often involved conditioned responses (e.g., T-bills, commercial paper *under 90 days*) and c-Rater needed these broken out or combined
- Concept development was time-consuming and nearly boundless
  - Closure on acceptable response set was nearly impossible
  - Concepts had to accommodate the case of a candidate giving a correct response followed by information indicating the response was not truly understood

72

Findings

- Complex sentence structure of responses and software limitations for human scoring input made some scoring decisions difficult (i.e., those incorporating two concepts in the same sentence, one with a verb, one in a phrase – c-Rater likes phrases with verbs)
- Candidates like to respond in lists, whereas c-Rater likes sentences – prompts would have to have been carefully designed to avoid this problem
- Atrocious spelling and grammar may have confounded c-Rater (and SMEs)
- Distracter analysis would be helpful in analyzing candidate misconceptions
- We might have excluded some obvious responses that provided little discrimination through the prompts

73

Findings

- Model creation required extensive computer time
  - Tens of different models tried to find ones that matched human scoring
  - Sometimes two days of computer running time required
  - Some of the models never worked
- Results were mixed
  - In some cases human-human agreement beat machine-human agreement performance
  - In some cases machine-human agreement beat human-human agreement
  - In some cases humans couldn't agree very well, making machine-human agreement impossible

74

Conclusions

- C-Rater works best with concepts that are clear, concise, and constrained
  - Such items are likely to be recall or definitional
  - Not a good fit for simulated tasks aimed at higher order skills
  - Likely a good replacement for non-quantitative MCQ items with well-defined answer sets

75

Conclusions

- Cost of development and model preparation would not justify use in our examination for most simulations or MCQ
- Cost might be justified in specific instances such a listening items where concepts are more constrained

Conclusions

- C-Rater might be put to good use for programs desiring to test true recall (vs. recognition) of simple concepts, e.g.,
  - Science
  - History
- C-Rater is unlikely to work well in professional assessment where concepts are likely to result in a multiplicity of equally valid responses

Speaker Contact Information

- Richard N. DeVore, Ed.D.
- rdevore@aicpa.org

Chaitanya Ramineni, PhD, F. Jay Breyer, PhD
Educational Testing Service
John Mattar, PhD
AICPA

AUTOMATED SCORING OF PERFORMANCE TASKS

Outline

- Background
- E-rater®
  - Evaluation criteria
  - Prompt-specific vs. Generic models
- E-rater for AICPA
  - Operation & Maintenance
  - Research

Background

- Two constructed response (CR) items administered in each of three test sections*
  - one scored by e-rater and one pre-test
  - the purpose of the item is to assess writing ability in the context of a job-related accounting task.
    - The response must be on topic but the primary focus of scoring is on writing ability.
    - If a response is determined to be off topic it is given a score of zero.

* Exam format revised beginning January 2011, all CR items now administered in one section

Background

- Human score a subset of pre-test responses
  - Use as the basis for building new e-rater automated scoring models
  - Each CR prompt has a custom-built (prompt-specific) model
    - Sample Test Constructed Response Item 2011.docx

82

CLEAR 2011
Annual Conference

September 8-10
Pittsburgh, Pennsylvania

e-rater®

- State-of-the-art automated scoring of English language essays
- e-rater scoring is similar to or better than the agreement standard set by human grading
- Most widely used ETS automated scoring capability, with more than 20 clients representing educational, practice and high-stakes uses, including:
  - Criterion, SAT Online, TOEFL Practice Online, GRE® ScoreItNow!℠, ETS® Proficiency Profile
  - GRE ® and TOEFL ®, among others

83

CLEAR 2011
Annual Conference

September 8-10
Pittsburgh, Pennsylvania

e-rater model development process

Evaluate items and rubrics for use with e-rater

1. Collect human scores
2. Split the data into model build and evaluation sets
3. Compute scoring features from the model build set
4. Determine optimal weights of features in predicting human scores (regression) from the model build set
5. Validate against additional human-scored cases in the evaluation set

84

CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

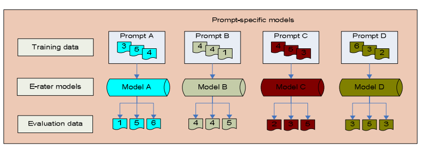
Evaluation criteria

- Construct relevance
- Empirical evidence of validity
  - Relationship to human scores
    - Agreement: Pearson *r* & wtd Kappa ≥ 0.70
    - Degradation: Reduction in *r* or wtd kappa from human-human agreement < 0.10
    - Scale: Difference in standardized mean scores < 0.15
  - Relationship to external criteria

85

CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

Model Types: Prompt-Specific

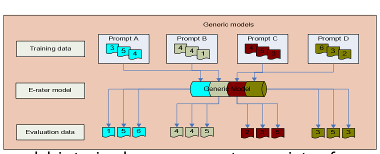


- Each model is trained on responses to a particular prompt
- Advantages:
  - Tailored to particular prompt characteristics
  - High agreement with human raters
- Disadvantages:
  - Higher demand for training data

86

CLEAR 2011 Annual Conference
September 8-10 Pittsburgh, Pennsylvania

Model Types: Generic



- A single model is trained on responses to a variety of prompts
- Potential advantages:
  - Smaller data set required for training.
  - Scoring standards the same across prompts.
- Disadvantages:
  - Features related to essay content cannot be used.
  - Differences between particular prompts are not accounted for.
  - Agreement with human raters is lower.

87

Operational use of e-rater (1)

- Responses for new pre-test items in each quarter are double scored by humans and the data are split into model build (~500 sample size)and evaluation set (all remaining responses)
- e-rater feature scores are computed on the model build set using the average human score as the criterion variable
- The feature scores are then applied to the evaluation set to evaluate e-rater model performance

88

Operational use of e-rater (2)

- e-rater models that meet the evaluation criteria are approved for operational use
- e-rater replaces human scoring for those items,
  - 5% responses, randomly selected, are rescored by humans for quality control purposes, and
  - Candidates close to the cut score (20-25%) are also rescored by humans
- Operational models are re-evaluated using new data when there are changes in the exam format (or upon client request)

89

Research with e-rater

- PS e-rater models have been approved for operational use for 78 prompts
- All CRs are human-scored using a common rubric, hence
  - Is a single overall (generic) model sufficient for all prompts?
- Research Plan: Using data for operational prompts, build and evaluate generic scoring model

90

Advantages of generic model

- More cost effective (than PS models) for large-scale assessments
  - Smaller sample sizes for model training
  - Consistent set of scoring criteria across prompts
- Streamline test development
  - Can create prompts that are similar and consistent in nature, by establishing a target model
  - Use same model to score new prompts

Results from 2009(1)

- Responses for 78 prompts with approved models were used
- Four generic models were built- overall and for each of the three test sections

	# of prompts	*N*	*Mean*	*SD*
Overall	78	38,848	2.67	0.93
Content area				
AUD	26	13,610	2.73	0.93
FAR	29	13,939	2.65	0.92
REG	23	11,299	2.60	0.95

Results (2)

- Evaluation sample results for PS and G models at the aggregate level

		Human1		*Auto*							
Prompt	N (avg)	Mean	SD	Mean	SD	% agree	% adj agree	kappa	wtd kappa	corr	Std diff
PS	470	2.66	0.92	2.66	0.91	64	99	0.49	0.75	0.76	0.01
All	498	2.66	0.92	2.66	0.83	60	98	0.42	0.70	0.73	-0.01
AUD	523	2.75	0.92	2.74	0.91	59	98	0.42	0.71	0.73	0.00
FAR	481	2.65	0.92	2.64	0.79	59	98	0.41	0.69	0.73	-0.01
REG	491	2.59	0.94	2.59	0.81	62	99	0.46	0.73	0.75	-0.01

Results (3)

- Flagging results at the prompt level

Model	*N*	*Wtd kappa flag*	*Correlation flag*	*Std diff flag*	*Total # of prompts flagged*
PS	78	6	6	0	6 (7%)
Overall	78	29	14	44	47 (60%)
AUD	26	11	10	11	15 (58%)
FAR	29	17	5	18	21 (72%)
REG	23	6	1	9	12 (52%)

94

Results from 2010(1)

- Responses for 34 (of the 78) prompts approved for inclusion in the new exam format (for 2011) were used to build a single generic scoring model
- Evaluation sample results for PS and G models at the aggregate level

		Human1		*Auto*		*Human 1 by Auto*					
Prompt	N	Mean	SD	Mean	SD	Std diff	kappa	wtd kappa	% agree	% adj agree	corr
PS34	514	2.68	0.91	2.69	0.90	0.00	0.52	0.77	66	99	0.81
GN34	808	2.68	0.91	2.69	0.86	0.00	0.46	0.74	62	99	0.80

95

Results (2)

- Flagging results at the prompt level

Model	*N*	*Std diff flag*	*Wtd kappa flag*	*Correlation flag*	*Total # of prompts flagged*
PS34	34	0	1	1	2 (6%)
GN34	34	16	5	1	16 (47%)

96

Results

- Prompt-specific models outperformed all generic models
- The performance of e-rater generic models is satisfactory at the aggregate level, however, concerns at prompt level
- High proportion of prompts flagged as problematic under each type of generic model

97

Research plan for 2011

- New exam format, different testing conditions
- Using operational data from the new exam format, build and evaluate generic scoring model

98

Dongyang Li, PhD
Prometric

SOME COMMENTS

99