The Assessment of Clinical Reasoning in Medical Education: A Review of the History, Current Approaches and Future Directions

Clarence D. Kreiter, PhD
University of Tokyo, IRCME
University of Iowa, Carver College of Medicine

Three Broad Areas in Medical Education Assessment

- 1. Clinical Knowledge
- 2. Clinical Skills (technical procedures, communication, professionalism…….)
- 3. Clinical Reasoning* (CR)

Why is it important to acquire a valid and reliable measure of CR?

- Many examples in science where advances in the science required accurate measurement

- "By comparing accurate measurements with numerical predictions of a theory, we can gain confidence the theory is correct." (Keith Symon, 1964 – Mechanics)

Why is it important? (con'd)

- Currently many theories regarding CR and how theory should guide learning
- All agree CR should be taught and therefore assessed
- But we currently lack a validated measure.

Why is it important? (con'd)

- Education Implications
  - May be the most important competency
  - Licensure Testing
  - Formative testing
  - Facilitates congruence between educational objectives and assessments

Why is it important? (con'd)

- Research implications:
  - Relationship between CR and General Reasoning
    - Selection research – Cognitive Research
  - Educational efficiency
    - Effectiveness of educational interventions
      - Assessing curriculum design
        - e.g. PBL vs Standard

Importance of Defining CR

- Validity considerations requires a clear definition
- Fuzzy definitions have negatively impact previous measurement efforts
  - Clinical Decision Making (CDM)
  - Clinical Reasoning (CR)
- Both have important role to play – but not the same

Simple Definition of CR

- Many variations and complex models – however for measurement a simple behavioral definition can adequately capture and describe the CR cognitive activity in medicine.

- "Clinical reasoning is a cognitive activity that integrates information from a clinical encounter with an existing system of knowledge organization" (Kreiter & Bergus – 2008)

Validity and Relation to Definition

- The goals is to detect and quantify variation in the ability to think logically within clinical domain
- Clinical decision making (CDM) may be of paramount interest, but only minor assumption required to link CR with its impact on CDM
- This figure captures what I think is important in this definition and helps define the measurement task

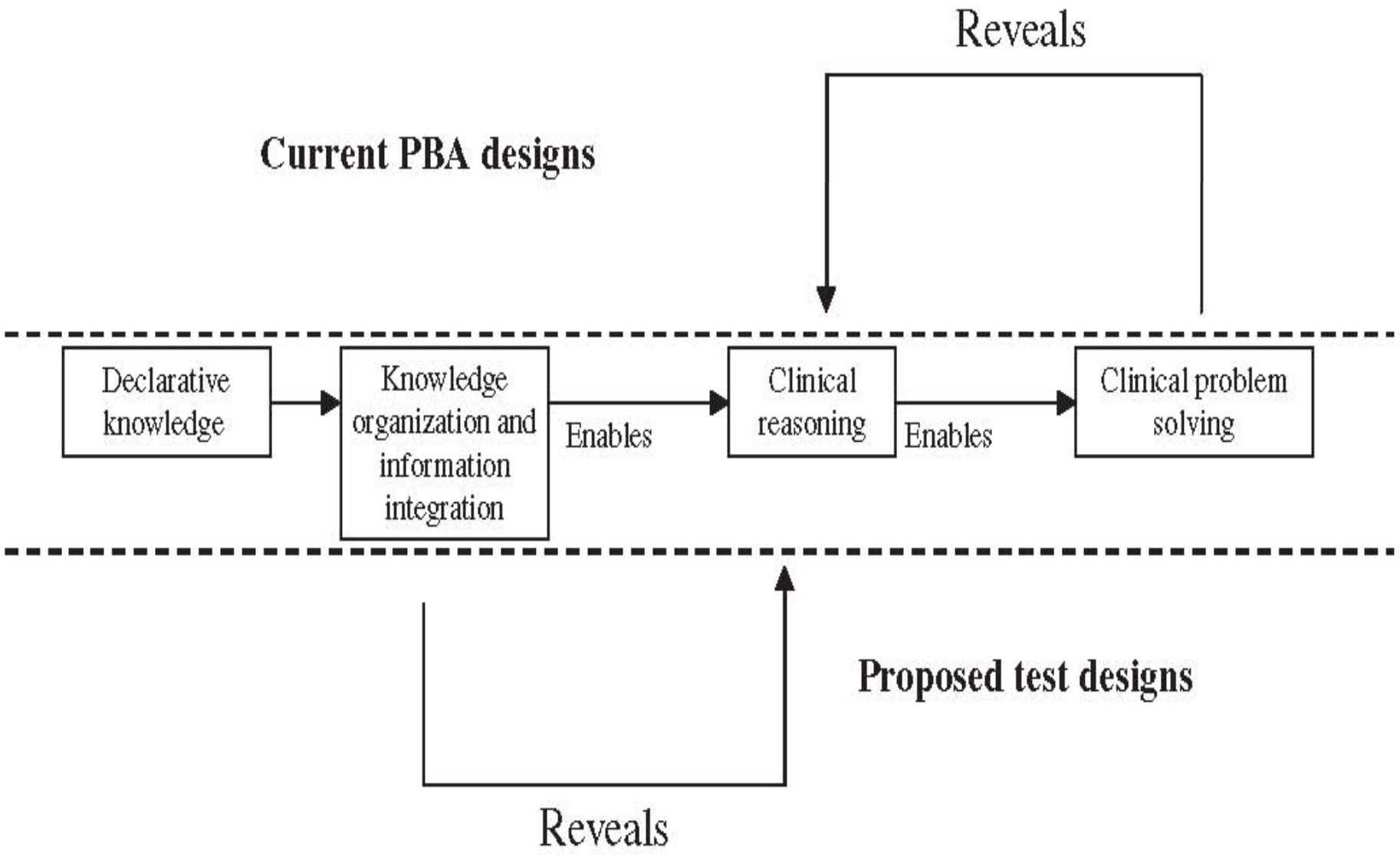


Figure 1 Important relationships for assessing clinical reasoning

Important validity consideration related to definition

- Integration of patient information into existing structures of knowledge (CR) enables clinical decision making (CDM)
- Definition of construct should guide test development and validity research – **theme of this talk**!
- But first some historical background

A Short History of CR Research

- 1970s – Several centers in North America began studying clinical reasoning
  - 'think aloud' method with SPs (MSU)
  - 'stimulated recall' method of recorded interaction (McMaster)
  - Case specificity observed (Elstein et al 1978 – 'Medical Problem Solving: An Analysis of Clinical Reasoning')
    - Measurement implications
    - Cognitive implications (Content Specificity)
    - Conclusion – no general ability

CR Research History (Con'd)

- 1970s - Case Specificity – profound impact
  - 'The low correlation between case performance suggests that intra-individual consistency on any of the dependent variables is weak and that scores on the variables are influenced much more by the structure of the problem and individual clinician's understanding of its task demand than by consistency in individual problem-solving style' (p 85 – Elstein et al.)

CR Research History (Con'd)

- 1970s - Case Specificity (an example where measurement influenced theory – however perhaps incorrectly)
  - Elstein (1978) looked at correlations btw cases – however attenuation alternate explanation.
  - Then Generalizability theory used – person by case interaction variance very high however interpretation of variance components questionable (Kreiter & Bergus 2007)

CR Research History (Con'd)

- Case Specificity
  - G studies in medical education generally concluded that inconsistency in CR performance is due to unique reasoning challenges presented by cases.
  - however interpretation of variance components questionable (Kreiter & Bergus 2007)
  - Many have confounding of the PC variance component – residual and hidden facets
  - Occasion facets never modeled

CR Research History

- Case Specificity – Evidence for the Occasion facet
  - Norman et al. (1985) looked at correlation between same cases given twice – still very low correlation
  - Shavalson et al. (1993) examined:
    - Person X Rater X Task X Occasion
    - Small (pt) pc interaction

CR Research History (Con'd)

- Case Specificity
  - Still has measurement implications – you need lots of cases if you plan to use Performance Assessment for CR
  - However evidence does NOT support past interpretation of case specificity of CR. (not highly multidimensional with each case requiring a unique composite of CR abilities) (Kreiter & Bergus, 2007)

CR Research History (Con'd)

- Case Specificity – Good News – Bad News
  - Bad News – Measurement evidence does not support Case Specificity of clinical reasoning – (40 yrs of research?)
  - Good News - Maybe we can replicate psychology's success in assessing reasoning

Success of Measuring Reasoning Construct in Psychology

- General Reasoning assessed with a high degree of accuracy (Wechsler Adult Intelligence Scale – WAIS - Matarrazzo, J.D. et al.)
  - Excellent accuracy / reliability (.85-.90) – strong validity evidence

Can we do the same for CR ability?

Comparing General Reasoning (GR) and Clinical Reasoning (CR) Assessments

- GR – Context-free and independent of knowledge
- CR – Context-bound and knowledge intensive
- Implies a valid CR test will be population-specific with assumptions related to educational level.

CR Research History (Con'd)

- 1980s - modeled on memory and expertise literature - Chess master example (Simon and Chase 1973)
  - Memory focus of research
  - Largely failed to replicate

- Memory of clinical encounter unlikely to be a CR measure

CR Research History (Con'd)

- 1990-2000s - Research on knowledge representation and structure
  - Knowledge representation ideas borrowed from cognitive psychology
    - Illness script - schemes
    - Exemplars derived form experience
    - Elaborated knowledge

    (Schmidt et al.) (Bordage & Lemieux)

Lessons from Research

- Experts display higher levels of knowledge organization
- Structure of knowledge organization idiosyncratic – multiple valid ways to structure
- Measures of exact knowledge structure not likely to yield valid measure since no gold standard for scoring
- Returning to Validity Model these lessons again highlight importance of measuring Step 2

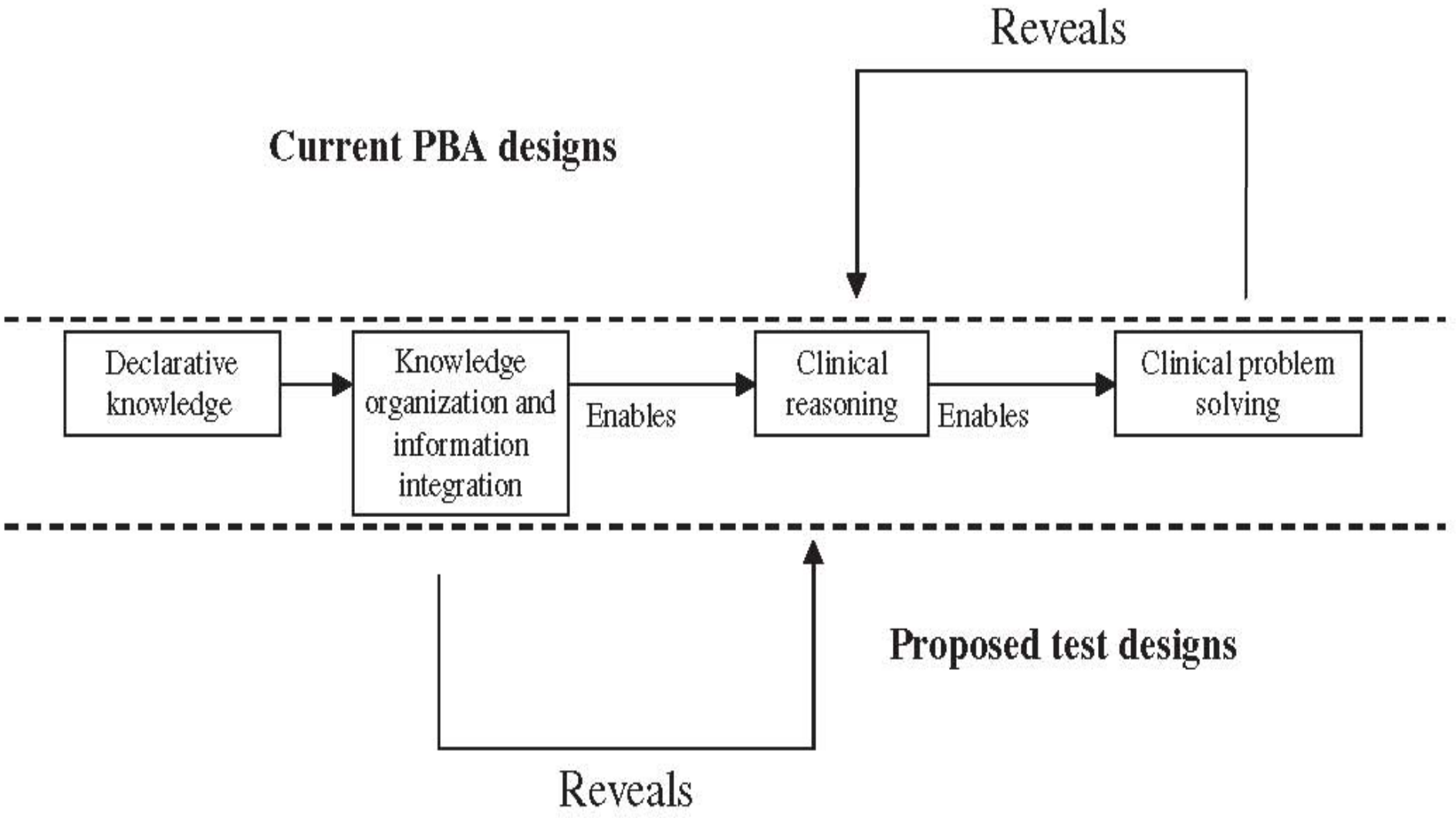


Figure 1 Important relationships for assessing clinical reasoning

Additional Lessons from CR Research

- Assessments are influenced by research
  - Theory influences measures and vice versa
  - Research advances require valid measures
  - CR research requires sound methodologies and measures

History of CR Assessment

- Simulation and Performance-Based Assessment (PBA)
  - Since clinical decision making (CDM) the outcome of CR, it seemed like the natural first step to simulate clinical problem environments and measure problem solving or CDM. But major problems……

History of CR Assessment

- Simulation and PBA
  - Simulations initially utilized simple paper and pencil methods
    - Patient Management Problems – Diag. Management Problems – Over 20 different kinds of PMPs reported in the literature – Used on certifications exams in 70s
      - Collect history, conduct physical exam, investigate, Dx, Treatment
      - Data collected sequentially – obscured until selected
      - Score based on data choices – pathways

History of CR Assessments

Scoring of PMPs, computer simulations (CCS), SP formats

- Optimal pathway
- Expert panel
- Expert use different pathways (i.e. low consensus)

History of CR Assessment

- Scoring evidence related to PMP, computer simulations (CCS), SP
  - Scoring methods reduce to simple measures of thoroughness
  - Unrelated to diagnostic accuracy
  - Inversely related to expertise (experts take shortcuts) (experts not better than less experienced)
  - Low case correlations (low reliability)

Lesson / Conclusion

- For a multitude of reasons, simulation data that is designed to summarize history, diagnosis, information gathering, and treatments are unlikely to yield a valid measure of CR.

Assessing Knowledge Organization and Structure

- Key Feature Items
- SCT
- Path Diagrams
- Case Vignettes
- Highly Structured Simulations

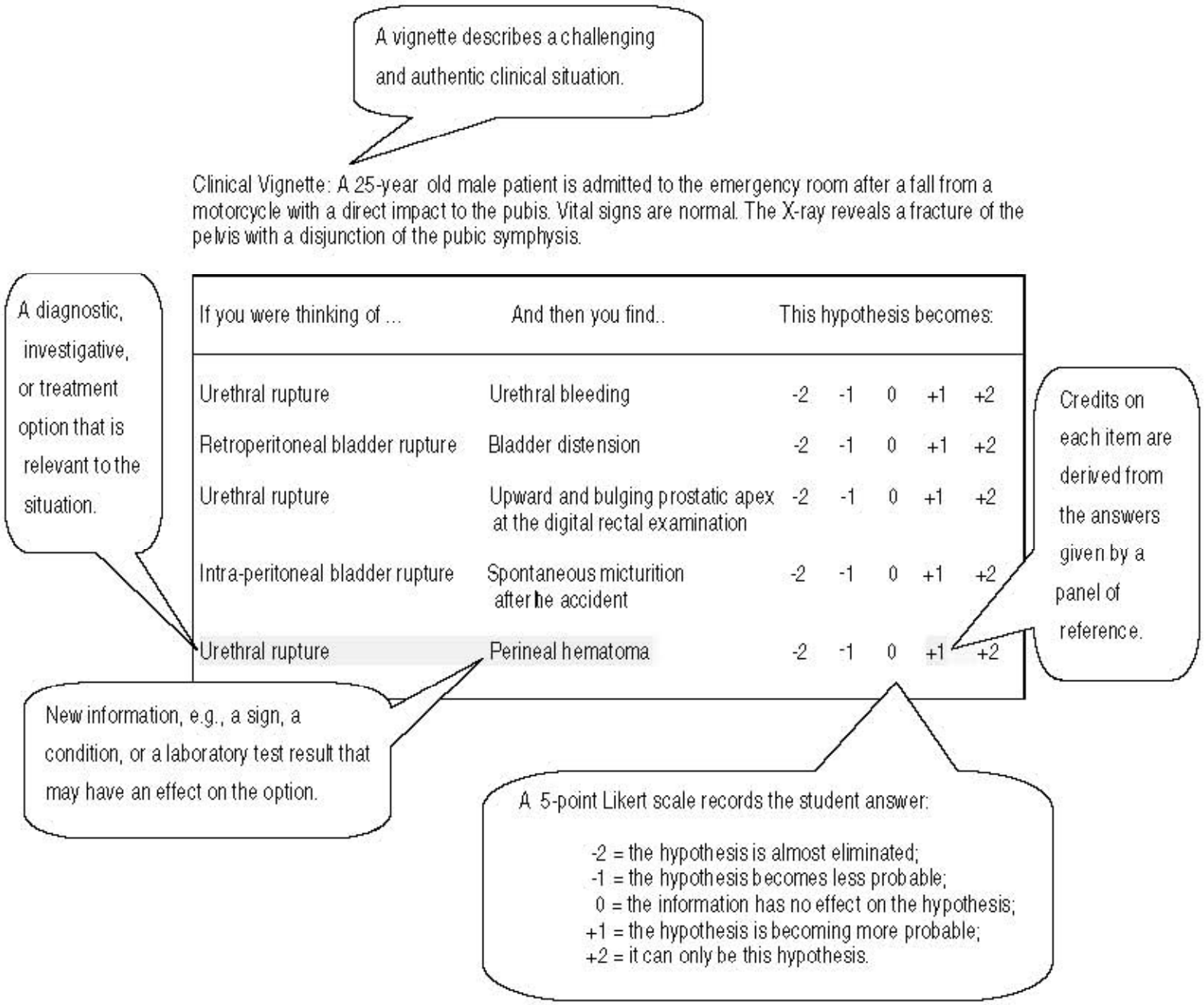
Key Feature Items

- Features*
  - Vignette
  - Selected response (short menu), or
  - Write-in response
  - May help guide item writers toward higher level cognitive questions
  - Used on Canadian licensure exams
  - Efficient
  - However Key Feature items can be rewritten as a MCQ formats using a clinical vignette with or without extended list responses format
  - Key Feature and MCQ true score correlation ~ .80 (CI includes 1.0)**

(Page, Bordage, Allen, 1995) * (Fischer, Kopp, Holzer, Ruderich, Junger, 2005)**

SCT

- Script Concordance Item (SCT)
  - Description and Example



The Test Format

Clinical Vignette: Joyce, 20 years old, is consulting at your office for a "vaginal discharge" she has been experiencing for the past week. She has had a new sexual partner for the past three months and she is worried about getting a sexually transmitted disease.

If You Were Thinking of (Infection)	And Then the Patient Reports or You Find on Clinical Examination	This Hypothesis Becomes				
Yeast	She had a sexually transmitted disease a few years ago	–2	–1	0	+1	+2
Chlamydia	She is taking a contraceptive pill	–2	–1	0	+1	+2
Herpes	She has an itchy vulvae	–2	–1	0	+1	+2
Herpes	She has dysuria	–2	–1	0	+1	+2
Yeast	Her discharge is greenish and itchy	–2	–1	0	+1	+2

Note: –2 = ruled out or almost ruled out; –1 = less probable; 0 = neither less nor more probable; +1 = more probable; +2 = certain or almost certain.

Example of Items from the Diagnostic Section of a Test

Script Concordance Testing

- Characteristics
  - No correct answer
  - May assess meaning of knowledge embedded in a clinical problem
  - Easy to write items
  - Requires Expert Panel
  - High Reliability
  - Construct = Reason Scripts
  - Much research on reliability and validity
  - May ask important question related to probabilities

SCT - Drawbacks

- Philosophical problem with no correct answer
- Correct answer scoring works as well
- Lack of consensus may be a result of scaling and confusion related to question meaning

(Bland, Kreiter & Gordon 2005)

SCT - Drawbacks

- Defined construct very questionable
- Exact task required by item poorly considered

Definition – ‘reasoning in the context of uncertainty’

SCT Drawbacks

- $S_1 > S_2 > P_1 > S_3 > P_2 > S_4 > \{P_2 - P_1\}$
- $P(D \mid T) = P(T|D) * P(D) / P(T)$
- SCT question asks about relation between

  P_1 and P_2 / $\{P_2 - P_1\}$

  $P(D)$ and $P(D \mid T)$ / $\{ P(D \mid T) - P(D) \}$
  - Why not ask about impact of T (S_3) ?

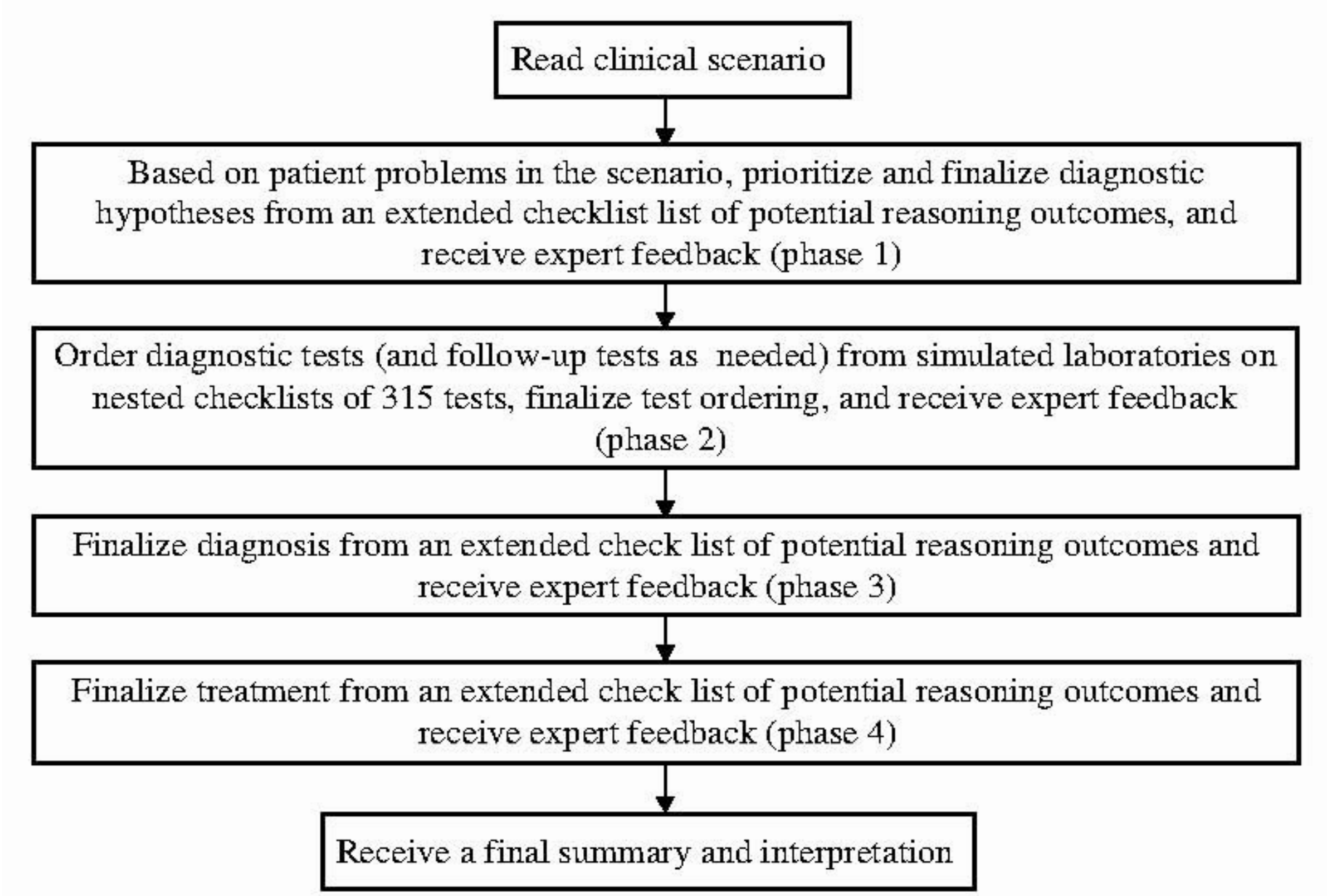
(Kreiter, 2011)

SCT

- All is not lost!!!
  - With simple changes item shows great promise (easy to write, reliable, probabilistic reasoning)
    - Changes Required
      - Correct answer scoring
      - Redefine construct

High Structure Simulations

- High structure may allow a method of overcoming scoring PBA scoring difficulties
- Areas where medicine has moved to computer may allow PBAs to function naturally as CR tests.
- LabCAPS
- (Kreiter, Haugen…McGaghie, Dee, 2010)



Flow chart of simulated patient workup.

Case # 11
Unit: Level 2
Patient Age: 63
Patient Sex: M

Clinical vignette

Progress notes
append / view

Prioritize differential diagnoses

Select tests

Submit order for selected tests

Information/Resources

Clinical Laboratory | Procedure-based Tests | Radiology/Nuclear Medicine

Blood Center
Chemistry
Hematology
Hemostasis
Immunology
Microbiology
Molecular Pathology
Urinalysis

Blood Center

- ☐ Antibody Screen (Indirect Coombs Test)
- ☐ Antibody Titration (IgM+IgG)
- ☐ Blood Type (ABO & Rh)
- ☐ Cold Agglutinin Titer
- ☐ Crossmatch
- ☐ Direct Coomb's Test (Direct Antiglobulin Test)
- ☐ Donath-Landsteiner Test
- ☐ Fetal Hemoglobin Screen
- ☐ HLA Antibody Detection Assay
- ☐ HLA Class I Typing
- ☐ Kleihauer-Betke test
- ☐ Platelet Antibody Screen Test

Chemistry

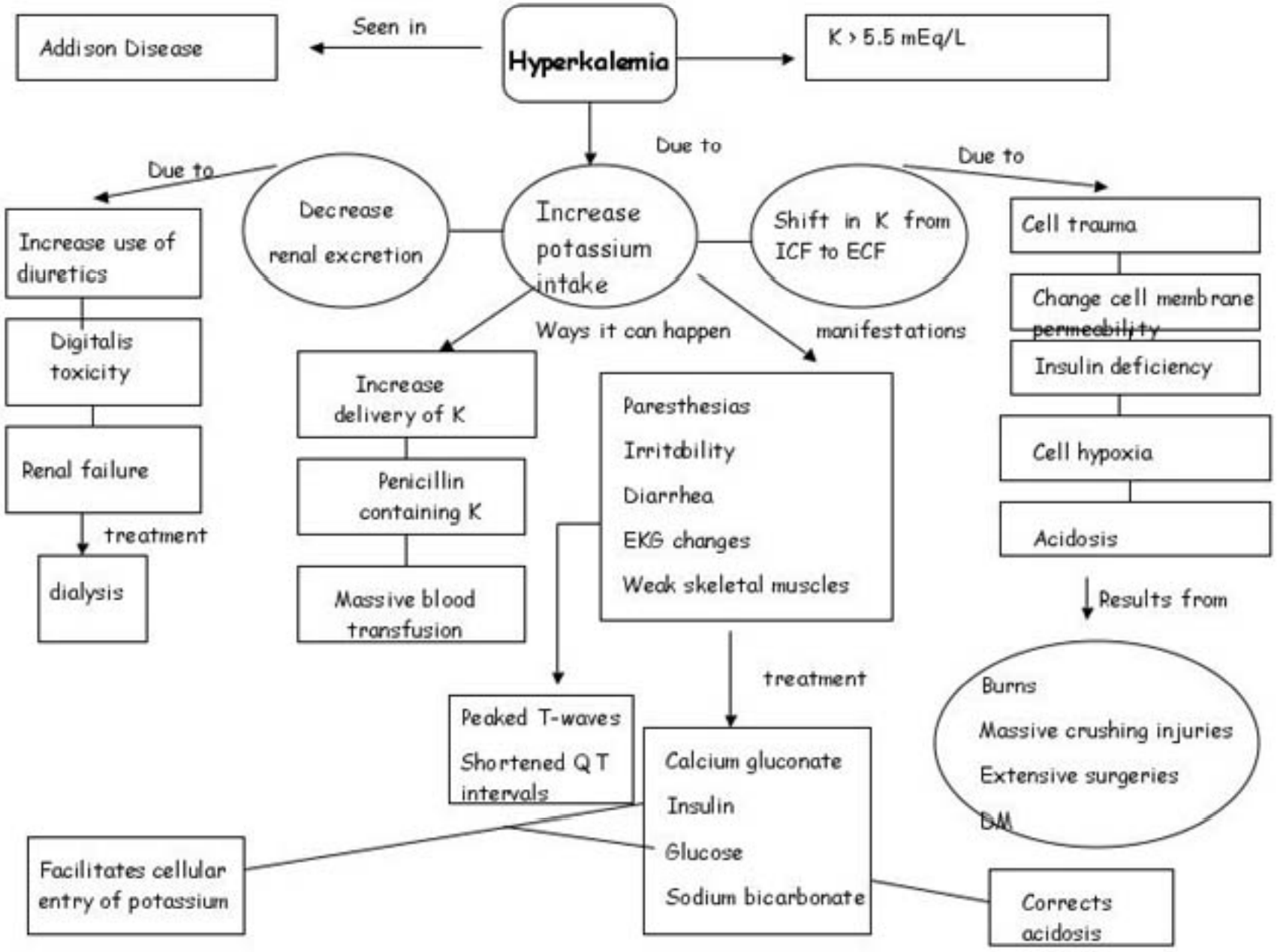
- ☐ Acid Phosphatase. Total
- ☐ Alanine Aminotransferase (ALT)
- ☐ Albumin
- ☐ Alpha Fetoprotein (nonpregnant)
- ☐ Ammonia
- ☐ Creatinine Kinase (CK) MB Isoenzyme (initial/baseline)
- ☐ D-Xylose. Urine
- ☐ Fat: Fecal Quanitative
- ☐ Ferritin. serum
- ☐ Folate. red cell
- ☐ Phosphatase. Alkaline
- ☐ Phosphorus
- ☐ Potassium
- ☐ Prostate Specific Antigen (PSA)
- ☑ Prostate Specific Antigen (PSA). free (includes total)

Example of screen used to order tests.

Path Diagrams / Concept Map

- Raters versus Objective scoring
- Idiosyncratic knowledge structures
- Interesting developments……

Example of Concept Map / Path Diagram



Case Vignettes with MCQ

- NBME
- Great Reliability
- Master items writers

(NBME – item writing guide)

An unresponsive 58-year-old woman is brought to the emergency department after collapsing at a local shopping mall. Her family reports that she felt well that morning but developed a headache that progressively worsened while she was shopping. She has had hypertension and atrial fibrillation and is taking an antihypertensive medication and an oral anticoagulant. Her blood pressure is 220/130 mm Hg and her respiratory pattern is one of apnea alternating with hyperpnea. She responds only to noxious stimuli with extensor posturing involving the right arm and leg. Fundoscopic examination reveals papilledema involving the left optic disc. Pupils are 3.0/7.0 (R/L) with no reaction to light on the left. There is a left gaze preference. There is diffuse hyperreflexia (R > L) and Babinski's sign is present bilaterally.

1. The dilated, unreactive left pupil is most consistent with injury to the left
   - A. optic nerve
   - B. optic tract
   - *C. oculomotor nerve
   - D. lateral geniculate nucleus
   - E. superior colliculus

2. The extensor posturing on the right is most consistent with injury to the left
   - A. telencephalon
   - B. diencephalon
   - *C. midbrain
   - D. pons
   - E. medulla

3. Her respiratory pattern is best described as
   - A. normal
   - *B. Cheyne-Stokes
   - C. central neurogenic hyperventilation
   - D. apneustic
   - E. ataxic

4. Which of the following herniation syndromes is most consistent with her clinical presentation?
   - A. Cingulate gyrus beneath the falx
   - *B. Temporal lobe uncus across the tentorium
   - C. Diencephalon through the tentorial notch
   - D. Brain stem through the tentorial notch
   - E. Cerebellar tonsils through the foramen magnum

Understandings Regarding CR Assessments

- Primary goal – test for : a command of substantive knowledge
- Logic tells us the ability to retain and express this knowledge is verbally mediated and is the basis of CR.
- This verbal knowledge can be communicated and stored in written format
- Verbal knowledge can be manipulated in the process of reasoning.

Understanding / Conclusions

- CR is embodied in the structure of verbal knowledge
  - can be listed as concepts and their inter-relationships.
  - Efforts to remove verbal knowledge from CR assessment misguided
    - To cite an exception in words allows us to add it to the list of important verbal knowledge
- CR as verbal knowledge can be tested with written formats.
- CR tests should target the verbal knowledge that represents the propositions and relationships that are the foundation medical clinical reasoning

References

- Matarrazzo, J.D., Carmody, T.P., Jacobs, L.D. Test-retest reliability and stability of the WAIS: A literature review with implication for clinical practice. Journal of Clinical Neuropsychology, 1980, 2(2) pp89-105.

- Elstein, A.S., Shulman, L.S., Sprafka, S.A. *Medical Problem solving: An Analysis of Clinical Reasoning.* Cambridge, MA: Harvard University Press, 1978.

- Schmidt, H.G., Norman, G.R., Boshuizen, H.P.A. A cognitive perspective on medical expertise: Theory Implications. Academic Medicine, 1990,65 pp611-21

- Kreiter, C.D., Bergus, G.R. Case Specificity: Empirical phenomenon or measurement artifact? Teaching and Learning in Medicine, 2007,19(4) pp378-381

- Norman, G.R., Tugwell, P., Feightner, J.W., Muzzin, L.J., Jacoby, L.L. Knowledge and clinical problem solving. Medical Education, 1985;19:344-56

- Shavelson, R.J., Baxter, G.P., Gao X. Sampling variability of performance assessments. Journal of Educational Measurement, 1993;30:215-32

References

- Bordage, G., Lemieux, M. Semantic structures and diagnostic thinking of experts and novices. Academic Medicine, *1991, 66,S70-S72.*

- Page, G., Bordage, G., Allen, T. Developing key-feature problems and examinations to assess clinical decision-making skills. Academic Medicine, 1995,70(3),194-201.

- Fischer, M.R., Kopp, V., Holzer, M., Ruderich, F., Junger, J. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Medical Teacher, 2005.

- Bland, A.C., Kreiter, C.D., Gordon, J.A. The psychometric properties of five scoring methods applied to the script concordance test. Academic Medicine, 2005,80(4),395-399.

- Kreiter, C.D. The response process validity of a script concordance test item. Advances in Health Science Education, - Published online 01 October 2011 – in advance of print

- Kreiter, C.D., Haugen, T., Leaven, T., Goerdt, C., Rosenthal, N., McGaghie, W.C., Dee, F. A report on the piloting of a novel computer-based medical case simulation for teaching and formative assessment of diagnostic laboratory testing. Medical Education Online, 2011, 16:5646,meov16.

Questions, questions……

- Questions?