Session: T232 – Salon B
Multi-Modal Summative Evaluation 2.0 – Improvement, Outcomes and Limitations

SOMU CHATTERJEE, MPH, M.D.
ANDREW R. WYANT, M. Div., M.D.

PAEA EDUCATION FORUM 2015
November 11–15 • Washington, DC

MULTI-MODAL SUMMATIVE EVALUATION 2.0 – IMPROVEMENT, OUTCOMES AND LIMITATIONS

SESSION# T232 / SALON B

ANDREW R. WYANT, M.DIV, M.D.
SOMU CHATTERJEE, MPH, M.D.

PRESENTATION OBJECTIVES

- Describe the process of multi-modal summative evaluation
- Compare and contrast the new methodology (2015) with previous yrs.
- Explain the rationale for changes in the methodology in the year 2015
- Describe how the data characterizes the process of multi-modal testing
- Describe what the data tells us about performance of students
- Summarize the effectiveness of the multimodal testing
- Limitations and future directions

BACKGROUND

ARC-PA / C3.04 – UPDATED 09/2014

- The program ***must conduct and document a summative evaluation of each student*** *within the final four months of the program to verify that each student is prepared to enter clinical practice.*
- ANNOTATION: Evaluation products designed primarily for individual student self-assessment, such as *PACKRAT are not to be used by programs to fulfill the summative evaluation of students within the final four months of the program. The ARC-PA expects that a program demonstrating compliance with the Standards will incorporate evaluation instrument/s that correlates with the didactic and clinical components of the program's curriculum and that measures* ***if the learner has the knowledge, interpersonal skills, patient care skills and professionalism*** *required to enter clinical practice.*

CORE COMPETENCIES

- Medical Knowledge
- Interpersonal & Communication Skills
- Patient Care
- Professionalism
- Practice-based Learning & Improvement
- Systems-based Practice

INTERPRETATION: ARC-PA REQUIREMENT & CORE COMPETENCIES INTEGRATION

- **Requirements of a *true* Summative Assessment:**

- Instrument must evaluate both didactic & clinical aspects of curriculum (C3.04)
- Metric evaluation of: (C3.04)
  - Medical Knowledge
  - Interpersonal / Communication Skills
  - Patient Care Skills
  - Professionalism
- Commensurate with entering clinical practice (C3.04)

FORMATIVE VS. SUMMATIVE

- **Formative Assessment:**
  - Low stakes
  - Identify students strengths / weaknesses
  - Help faculty recognize student's struggles
- **Summative Assessment:**
  - High Stakes
  - Evaluate student learning at the end of a unit
  - Compared to a *benchmark or standard*

LITERATURE REVIEW: LOOKING FOR A BENCHMARK?

In the field of PA education, with the exception of the PACKRAT, there is a lack of validated formative and summative instruments that can accurately predict future PANCE performance……. S. Massey, et al, JPAE 2013 vol24 No1

Logistic regression analyses showed a significant relationship between PACKRAT scores and PANCE (p < .001, cc 0.67),,, PACKRAT score >55%, with sensitivity of 77%.. predicting a strong correlation between PACKRAT & PANCE….. Cody, et al, PPAE 2004

Retrospective analysis of multiple potential academic and predictive factors showed that a year one GPA of < 3.0, or a Summative Exam score < 67% associated with an increased risk of PANCE failure…Ennulat C, et al, JPAE 2011, vol22 No1

Pilot study using a 360 MCQ-formative, and 700 MCQ- summative exam revealed correlative data with PANCE performance… S Massey, et al, JPAE 2013 vol24 No1

LITERATURE REVIEW: LOOKING FOR A BENCHMARK?

Objective structured clinical examination (OSCE) utilizing simulation as a predictor of clinical performance in the final year of training, is both valid and reliable for assessment of competence…Petrusa ER, et al, Arch Internal Med …1990;150:573-77

In the field of PA education, with the exception of the PACKRAT, there is a lack of validated formative and summative instruments that can accurately predict future PANCE performance….. Massey S, et al, JPAE 2013, Vol24 No1

THE PROCESS

MULTI-MODAL SUMMATIVE EXAM (MMSE)

- **Part A**
  - Simulated Test using a Standardized Patient
  - ✓Medical Knowledge
  - ✓Interpersonal & Communication Skills (verbal & written)
  - ✓Patient Care
  - ✓Professionalism
- **Part B**
  - Critique a relevant scientific article with essay responses
  - ✓Written Communication
  - ✓Practice based Learning & Improvement
  - ✓Systems-based Practice

EVALUATION OF THE STUDENT

- Evaluated by Standardized Patient (SP) – points based checklist
- Evaluated by Faculty – points based checklist
- Graded on SOAP Notes – rubric *(New in 2015)*
- 'Blink' Score *(New in 2015)*
- Scientific article with essay response graded by faculty *(New Rubric)*

CHANGES IN 2015 (2.0)

- Faculty were trained and oriented
- Standardized Patient training
- Cases were rewritten: (vetted and reviewed)
  - Tasks defined clearly
  - Outcomes defined clearly
- Rubrics redesigned or added
  - New Rubrics for faculty, for SP, SOAP note and the scientific paper
- 'Blink' score introduced

EVALUATION RUBRIC USED BY FACULTY

- **Presentation**
  - History taking
  - Physical Exam
  - Logic
  - Working plan
- **Clinical Problem solving**
  - Primary and Secondary diagnoses
  - Laboratory tests
  - Understanding the level of acuity
  - Assessment and plan
- **Patient Education**
  - Discusses diagnosis and plan of action with patient in layman terms

EVALUATION RUBRIC USED BY STANDARDIZED PATIENT

- Professional Introduction
  - Washed hands?
  - States level of training
  - Clear communication
- History
  - Addresses primary and related symptoms
  - Addresses symptoms related to Diff. Diagnoses
- Physical Exam
  - General
  - System specific
  - Other systems as necessary

‘BLINK’ SCORE

- Intuitive score given by faculty after seeing the student interact with patient
- Non verbal cues
- Rationale:
  - Gestalt (overall) score after observing the student perform in contrast to the score given before the start as done in other studies.*

*2009 Society for Academic Emergency Medicine (SAEM) Annual Meeting Abstracts. BLINK: Faculty First-Impressions of Emergency Medicine Residency Applicants Are Not Accurate. David Slattery, Mike Epter, Ross Berkeley, University of Nevada School of Medicine

DATA ANALYSIS

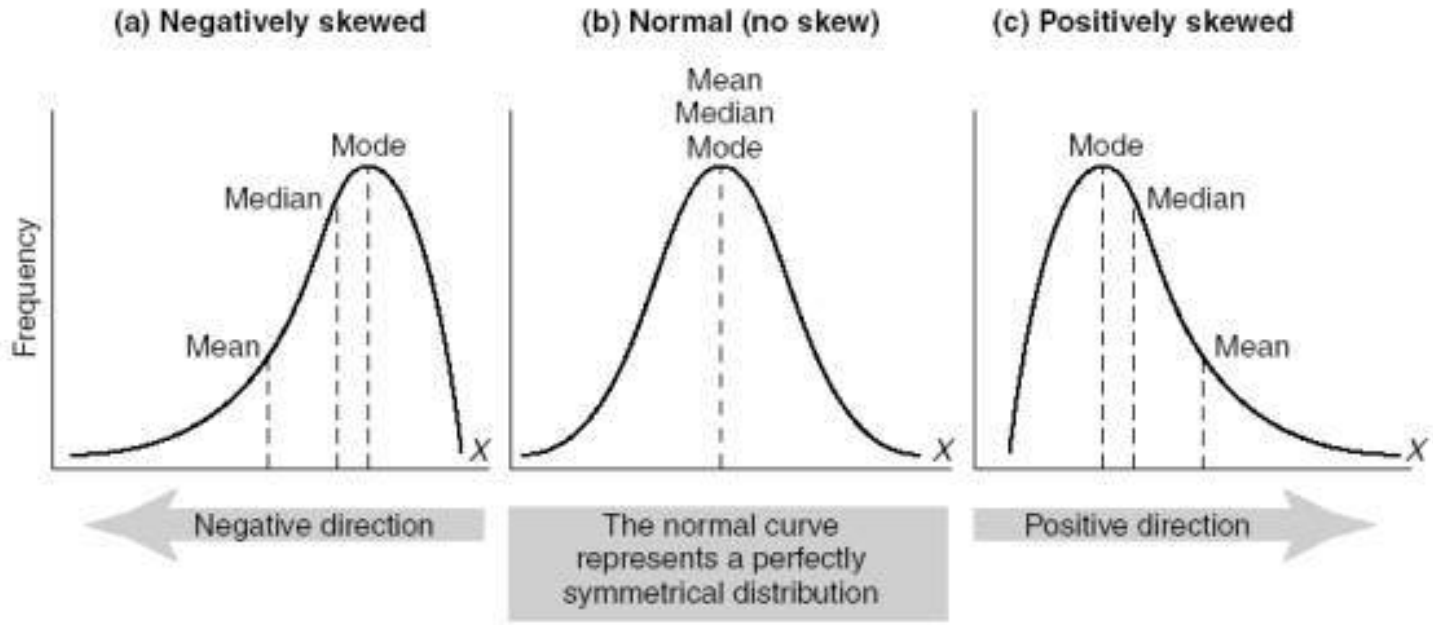
GOAL OF MMSE 2.0

- To design a summative evaluation that will do the following:
  - Provide consistency in grading
    - Types of data collected from the Multi-Modal exercise
    - Is there any inter-rater variability for
      - ✓All students
      - ✓Any subset of students?
    - Correlation between SP grading and Faculty grading
  - Provide strong correlation between success in MMSE and success in PANCE etc.

SUMMARY DESCRIPTIVE STATISTICS

- N=29 Students in the SP testing cohort
- Four main evaluation methods analyzed
  - SP checklist scores (passing score 60%)
  - Faculty checklist scores (passing score 60%)
  - Blink scores (out of 130 reported in %)
  - PANCE scores (as reported)
- Excluded
  - Students who did not opt for SP testing
  - Students who did not take the PANCE exam

SKEW

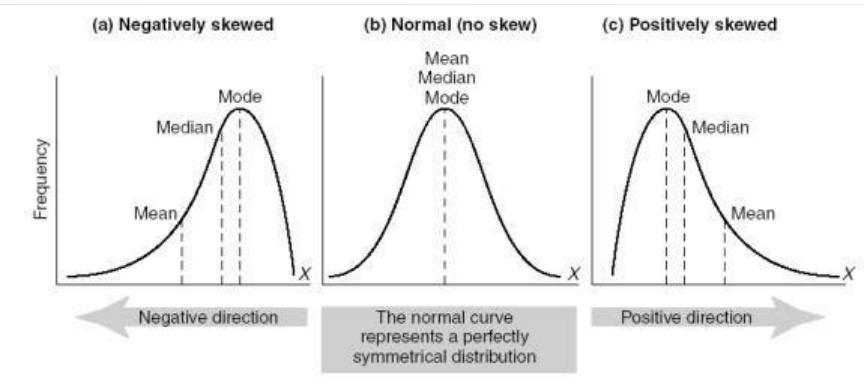


https://www.southalabama.edu

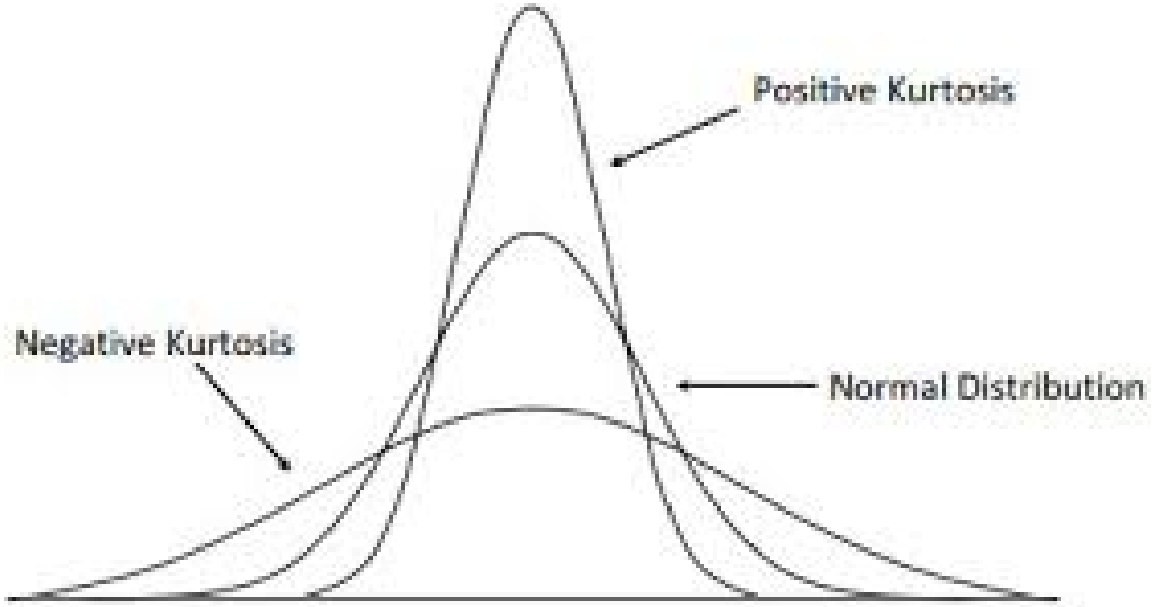
SKEW OF MAJOR GROUPS OF EVALUATIONS

	SP Checklist	Faculty Checklist	Blink score	PANCE score
Skewness	-0.132660204	-0.782213684	-1.591645934	0.138514079

- Skewness quantifies how symmetrical the distribution is.
- If the skewness is greater than 1.0 (or less than -1.0), the skewness is substantial and the distribution is far from symmetrical.
- An asymmetrical distribution with a long tail to the left (lower values) has a negative skew.
- SP, Faculty and PANCE scores are almost normal distribution, SP > Faculty ? Predispostion/Bias



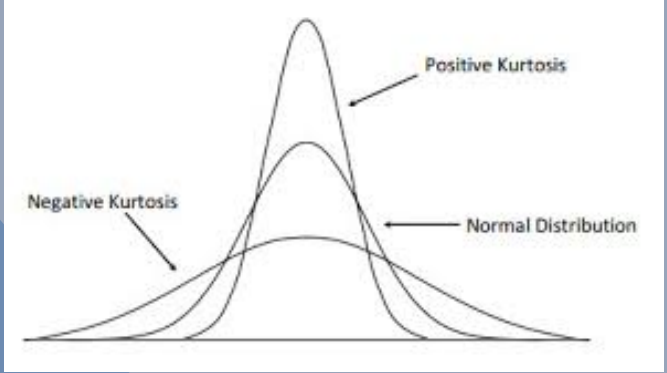
KURTOSIS


Positive Kurtosis
Negative Kurtosis
Normal Distribution

KURTOSIS OF MAJOR GROUPS OF EVALUATIONS

	SP Checklist	Faculty Checklist	Blink score	PANCE score
Kurtosis	-0.780686464	1.124953223	4.211957039	-0.378089438

- Kurtosis quantifies the variability of the scores
- < 3 is flat – more fat tails – more variable
- =3 is normal distribution.
- >3 is peaked – grouped tightly around the mean – tall peak.
- "Blink" score does not have normal distribution -? Putting more thought into it



CORRELATION COEFFICIENT BETWEEN EVALUATIONS

- Exactly –1. A perfect downhill (negative) linear relationship
- –0.70. A strong downhill (negative) linear relationship
- –0.50. A moderate downhill (negative) relationship
- –0.30. A weak downhill (negative) linear relationship
- 0. No linear relationship
- +0.30. A weak uphill (positive) linear relationship
- +0.50. A moderate uphill (positive) relationship
- +0.70. A strong uphill (positive) linear relationship
- Exactly +1. A perfect uphill (positive) linear relationship
- Consistency in faculty scoring

	Correlation r
SP & Faculty scores	-0.20
SP & PANCE	-0.10
Faculty scores & PANCE	0.23
BLINK score & PANCE	0.43
Total scores & PANCE	-0.03
BLINK score & SP score	-0.09
BLINK Score & Faculty scores	0.61

TESTING THE SCORES BETWEEN SP AND FACULTY

Null Hypothesis: There is no significant difference between the means of scores given by the SPs and the Faculty

Rationale:

- Same student
- Subjected to 2 instruments
- Similar to “Before” and “After” test

Paired t-test		
	SP Assessment Checklist	Faculty Checklist
Mean	36.76	23.45
Variance	28.26	3.02
Observations	29	29
Pearson Correlation	-0.202289065	
Hypothesized Mean Difference	0	
df	28	
t Stat	12.11	
P(T<=t) one-tail	5.98666E-13	
t Critical one-tail	1.70	
P(T<=t) two-tail	1.19733E-12	
t Critical two-tail	2.05	

Conclusion : They are significantly different

CORRELATION IN SUB-GROUPS (TOTAL SCORE VS. PANCE SCORES)

Correlation of scores of students within ±1SD	-0.037804105
Correlation of scores of students <-1SD	-0.526974769
Correlation of scores of students above >+1SD	0.997695853

- The SP + Faculty scores (>+1SD) are positively associated with higher scores (>+1SD) on PANCE
- SP + Faculty scoring (<-1SD) is not associated with PANCE scores at <-1SD

CALCULATING MEASURES OF VALIDITY AT THE UPPER END OF SCORES

	PANCE >+1SD		
SP+Faculty	Yes	No	Total
>+1SD	1	4	5
<+1SD	3	21	24
Total	4	25	29

	Estimated Value
Prevalence	0.137931
Sensitivity	0.25
Specificity	0.84

For any particular positive test result, the probability that it is:			
True Positive	0.2	0.01053	0.701208
False Positive	0.8	0.298792	0.98947
For any particular negative test result, the probability that it is:			
True Negative	0.875	0.66539	0.967147
False Negative	0.125	0.032853	0.33461

Specificity=Among all who score <+1SD in PANCE, the SP + Faculty scores <+1SD *will detect* those students 84% of the time.

NPV=Among all students who score <+1SD in Faculty +SP scores, 87.5% of the time they will score <+1SD in PANCE

CALCULATING MEASURES OF VALIDITY
LOWER END OF SCORES

	PANCE <-1SD		
SP+Faculty	Yes	No	Total
<-1SD	0	3	26
>-1SD	5	21	3
Total	5	24	29

	Estimated Value
Prevalence	0.172414
Sensitivity	0
Specificity	0.875

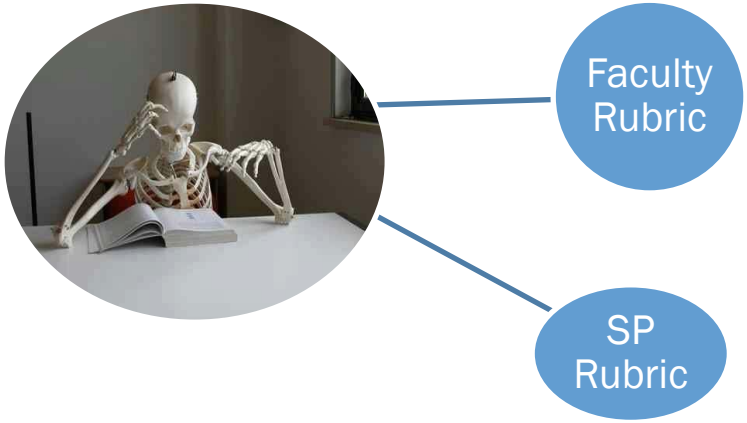
For any particular positive test result, the probability that it is:			
True Positive	0	0	0.690011
False Positive	1	0.309989	1
For any particular negative test result, the probability that it is:			
True Negative	0.807692	0.600168	0.926947
False Negative	0.192308	0.073053	0.399832

Specificity=Among all who score >-1SD in PANCE, the SP + Faculty scores >-1SD *will detect* those students 87.5% of the time.

NPV=Among all students who score >-1SD in Faculty +SP scores, 80.7% of the time they will score >-1SD in PANCE

INTER-RATER RELIABILITY – KAPPA COEFFICIENT

- Decided against doing it because of the following:
- On one student
- Testing with 2 different instruments
- Testing with 2 different observers handling the two different instruments
- Too much confounding with less data



Student	SP	Faculty
S1	SP1	F1
S2	SP2	F2
S3	Sp3	F3

CONCLUSIONS

SUMMARY

- Skew – SP > Faculty ? Predispostion/Bias
- Kurtosis – “Blink” score does not have normal distribution -? Putting more thought into it
- +ve Correlation
  - Faculty score & PANCE
  - Blink score and PANCE
  - Blink and Faculty scores - Consistency in faculty scoring

SUMMARY

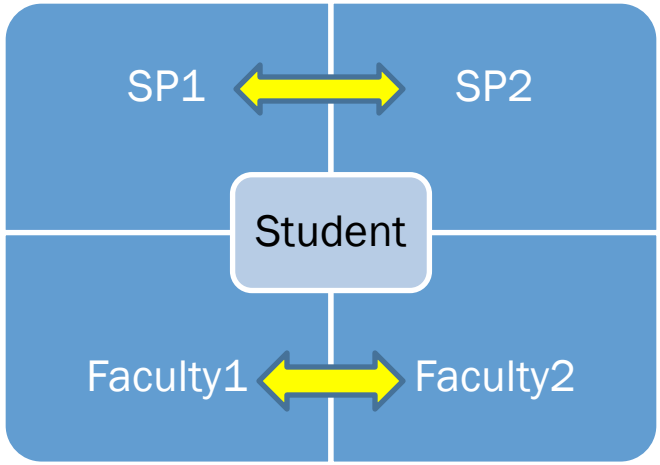
- Correlation in subgroups >+1SD, between +1SD and -1SD, <-1SD
  - The SP + Faculty scores are positively (and linearly) associated with scores >+1SD in PANCE
  - SP + Faculty scoring is negatively correlated with scores <-1SD on PANCE
  - No relationship for those within +1SD and -1SD
- SP and Faculty scores are significantly different (apples & oranges?)
- Data for validity
  - Specificity – 84% at the higher end <+1SD; 87.5% at the lower end >-1SD
  - NPV – 87.5% at the higher end < +1SD; 80.7% at the lower end >-1SD
  - Too many False positives/ Not sensitive at either end.

LESSONS LEARNT

- For improving internal validity we need to improve the grading mechanisms of SPs and Faculty
- Need to conduct a multivariate analysis to figure out which sections in the rubric is giving rise to most variations among the graders
- Address issues by way of training and address bias
- Cannot assess inter-rater variability until we have enough data points
- Have enough data points to have sensitivity and specificity to score above/below a certain threshold – difficult with only one cohort
- SP testing or Medical knowledge alone is not sufficient to be tested for PANCE certification??

FUTURE DIRECTIONS

- Changes – preserve the holistic format but switch the SPs and Faculty



- Get few more years of data
- Subgroup analysis
- Multivariate analysis on Rubric
- Better training of SPs and Faculty
- Address subjective /unconscious Biases

THANK YOU

Q?