Psychometric Analysis
IMG WBA Cohorts 2010-2012

Presenter - Associate Professor Curtis Lee

Special thanks to

Professor Balakrishnan R (Kichu) Nair AM
Director – Centre for Medical Professional Development

Professor Brian Jolly
Faculty of Health and Medicine
University of Newcastle

Analysis Demographics

Data Collected from Work Based Assessment program

- Mini-CEX x 12 stations
- CBDs (2 real cases and 5 paper cases)
- 360 data

Cohort Data

- 82 IMGs over 4 semesters

  2010:2 27 IMGs

  2011:1 22 IMGs

  2011:2 19 IMGs

  2012:1 14 IMGs

Data Collected for IMG pilot

Over 4 cohorts each IMG were assessed by:

12 Mini-CEX encounters (7 data points each)

7 Case Based assessments (7 data points each)

360°: on average 8 colleagues (23 data points)

on average 8 co-workers (23 data points)

on average 2 self reports (25 data points)

In the database:

3058 assessments

45 619 datapoints

Generalisability Theory

- Allows for the estimation of variance components associated with different examination conditions.

- In this case, for the Mini-CEX and the CBD, what *facet* is the greatest source of variability of the scores?

Mini-CEX

12 encounters classified as:

- Clinical Areas
  - Emergency, Medicine, Mental Health, O & G, Paediatrics or Surgery
- Domain
  - Counselling & Patient Education, History, History & Investigation, Investigations & Diagnostics, Management, Management & Counselling, Management & Prescriptions, and Physical Examination

Mini-CEX - Generalisibility study*

G-coefficient = 0.693

Facet	% Variance
IMG	2.3%
IMG x Cohort	2.6%
IMG x Area	14.5%
IMG x Domain	9.8%
IMG x Cohort x Area	9.4 %
IMG x Domain x Area	12.5%
IMG x Domain x Area x Cohort	0.1%
Residual	30.7%

Cohort = Semester (2010, 2011:1, 2011:2, 2012:1)

Area = Paediatrics, Mental Health, etc

Domain = PE, Management, History, etc.

*MINQUE rotation

Mini-CEX - Generalisibility study

- Traditionally, we would like to see the greatest variability by IMG
  - The assessment should be consistent and ONLY differentiate between those assessed. This was NOT the case for this dataset (only 2.3% of variance)
  - Average score of IMGs on the Mini-CEX was 5.8 with a standard deviation of 1.5
- Majority of accounted variance is by 'IMG x Area' and 'IMG x Domain x Area' (27%)
  - i.e. more variation if it is a paediatric case or a mental health case.
- Little variation from semester to semester (2.6%)

Mini-CEX - Generalisibility study

- What could this mean?
  - Perhaps IMGs are much better in some areas than others but have similar overall scores?
  - Perhaps items by area and by domain are of different levels of difficulty?
  - Perhaps the scoring by area and/or domain results in different scores?

Case Based Discussions (CBDs)

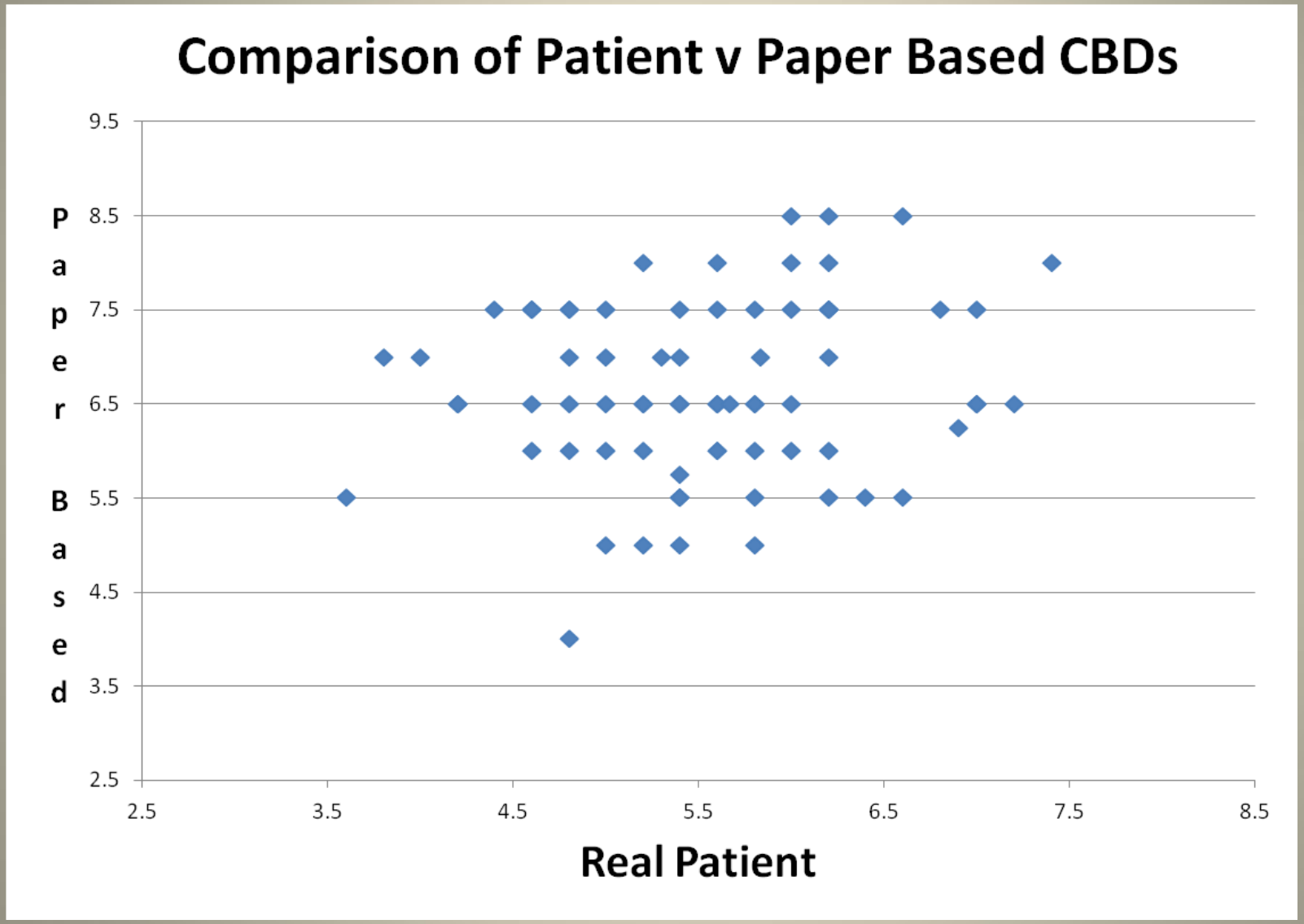
- Each IMG was administered 7 CBDs
  - 2 real patient encounters
  - 5 paper-based assessments (based on previous real cases)
- The cases
  - 6 Clinical Areas

    Emergency, Medicine, Mental Health, O & G, Paediatrics, and Surgery
  - Paper scenarios covered 5 of these per IMG.
- Scoring was based on content of the case – not a consistent rating scale like the Mini-CEX

Is there a difference between the Paper and Patient based CBDs?

- Using a paired t-test, scores from the Patient-cases were significantly higher than the Paper cases (Means of 6.7 vs 5.5 at $p < 0.05$)

- When correlating the scores IMGs received from their paper-based cases vs their patient-based assessments, the correlation was very low (Pearson's $r = 0.18$)


Comparison of Patient v Paper Based CBDs
Paper Based
9.5
8.5
7.5
6.5
5.5
4.5
3.5
2.5
2.5
3.5
4.5
5.5
6.5
7.5
8.5
Real Patient

Are there differences by content areas?

	Emerg	Medicine	MentalH	OandG	Paeds	Surgery	Real1	Real2
Emerg	1.00	0.16	0.20	0.46	0.20	0.33	0.31	**-0.05**
Medicine		1.00	0.21	0.09	0.03	0.05	0.07	**-0.04**
MentalH			1.00	0.34	0.17	0.12	0.20	0.05
OandG				1.00	0.19	0.04	0.26	**-0.04**
Paeds					1.00	0.27	0.18	0.01
Surgery						1.00	0.13	**-0.11**
Real1							1.00	0.12
Real2								1.00

CBD - Generalisibility study*

G-coefficient = 0.434

Facet	% Variance
IMG	9.8%
IMG x Type	0.0%
IMG x Domain	12.7%
IMG x Domain x Type	0.0%
Residual	77.5%

Type = Patient or Paper based CBD

Domain = Emergency, Medicine, etc.

*MINQUE rotation

Summary of analysis of CBDs

- Are the CBDs reliable?
  - Based on the CBDs alone, the classical reliability (as measured with Cronbach's Alpha r=0.566) is low. The Generalisability coefficient (0.436) is lower suggesting other factors may be contributing to the variance.

- What causes the inconsistency for the CBDs?
  - For the mini-CEX, the same scale is used for each encounter reducing the inter-station variability. This is not the case for CBDs as the cases are content based and the scores are likely more reflective of the difficulty of the case than an equivalent comparison between cases. This likely causes the model to show excessive variability due to the cases and little variability for other factors.

Summary of analysis of CBDs

(continued…)

- Why is the IMG variance in CBD so large compared with that in MiniCEX?
  - The reported variance is a percentage of the variance explained by the model. The variance for the IMGs is lower in the CBDs may be because the model doesn't account for some variable (ie something we didn't track) like communication ability or language or case grading consistency.

- Which is better: Patient or Paper based CBDs?
  - Based on the analysis, there is little difference in what each contributes to the model. There is greater variability between the content areas. Something else (like rating scales) have a greater impact.

Value of the CBD

Given this, what is the value of CBD?

- Van der Vleuten and others have stated that assessments should be valued for **validity, reliability, practicality and educational value.**
- CBDs may not have high reliability but their content validity is high and they provide a practical assessment of a broad range of content areas.
- They allow the review of a broad variety of real cases in a structured manner; different than PBL or chart stimulated recall.

Conclusions

- Patient based or paper based CBDs assess different aspects of competence and likely make the assessments more comprehensive. The reliability of the encounters are low but they provide content validity comprehensiveness.

- What is the purpose of these assessments?

Conclusions

- What is most important for assessments at this stage?
  - Are they formative? summative?
  - How defensible are they?
  - Comprehensiveness of assessment?
  - Statistical consistency?
  - Validity? Reliability?

- Based on these statistics, how can these assessment be improved?
  - More standardised assessments?
  - Rating scales?

Thank you

If you have questions or comments,
please contact Curtis Lee

curtis.lee@newcastle.edu.au