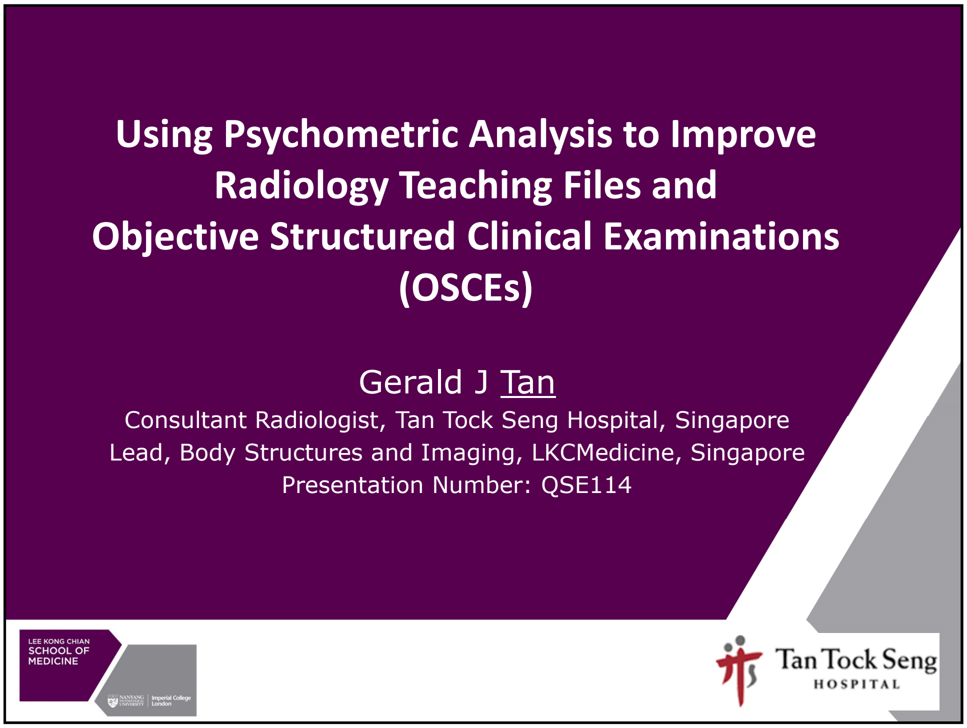
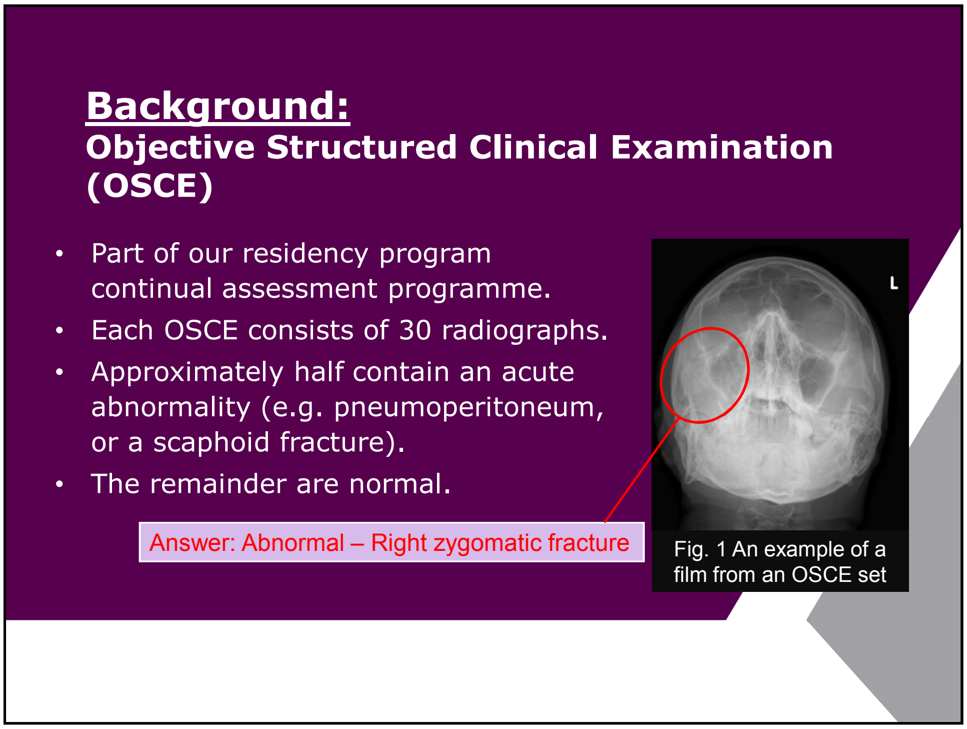


Using Psychometric Analysis to Improve Radiology Teaching Files and Objective Structured Clinical Examinations (OSCEs)

Gerald J Tan
Consultant Radiologist, Tan Tock Seng Hospital, Singapore
Lead, Body Structures and Imaging, LKCMedicine, Singapore

Presentation Number: QSE114



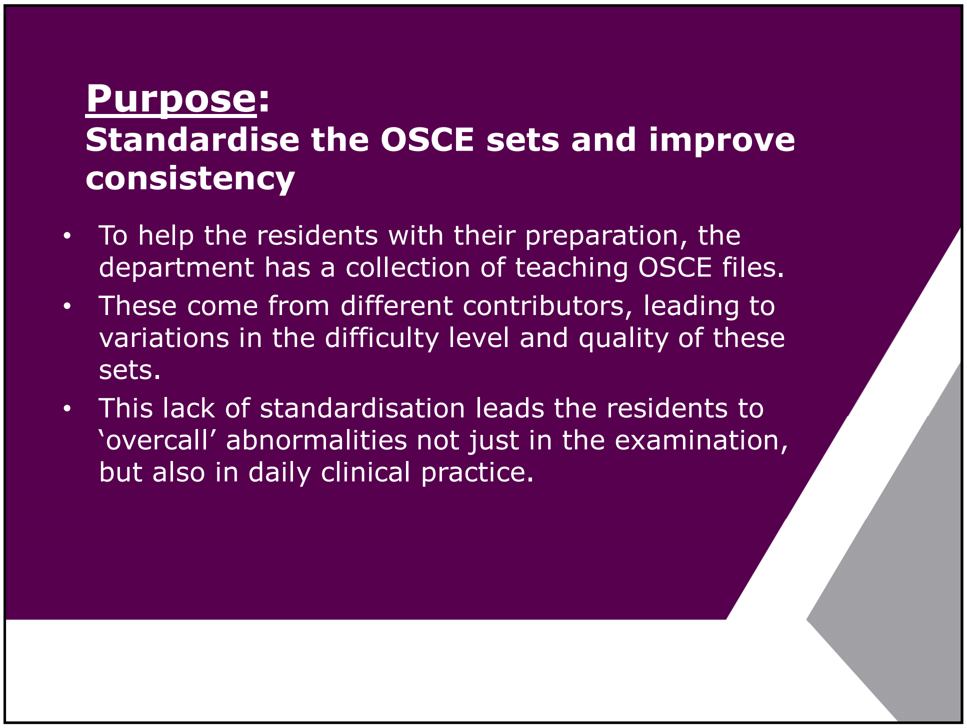
Background:

Objective Structured Clinical Examination (OSCE)

- Part of our residency program continual assessment programme.
- Each OSCE consists of 30 radiographs.
- Approximately half contain an acute abnormality (e.g. pneumoperitoneum, or a scaphoid fracture).
- The remainder are normal.

Answer: Abnormal – Right zygomatic fracture

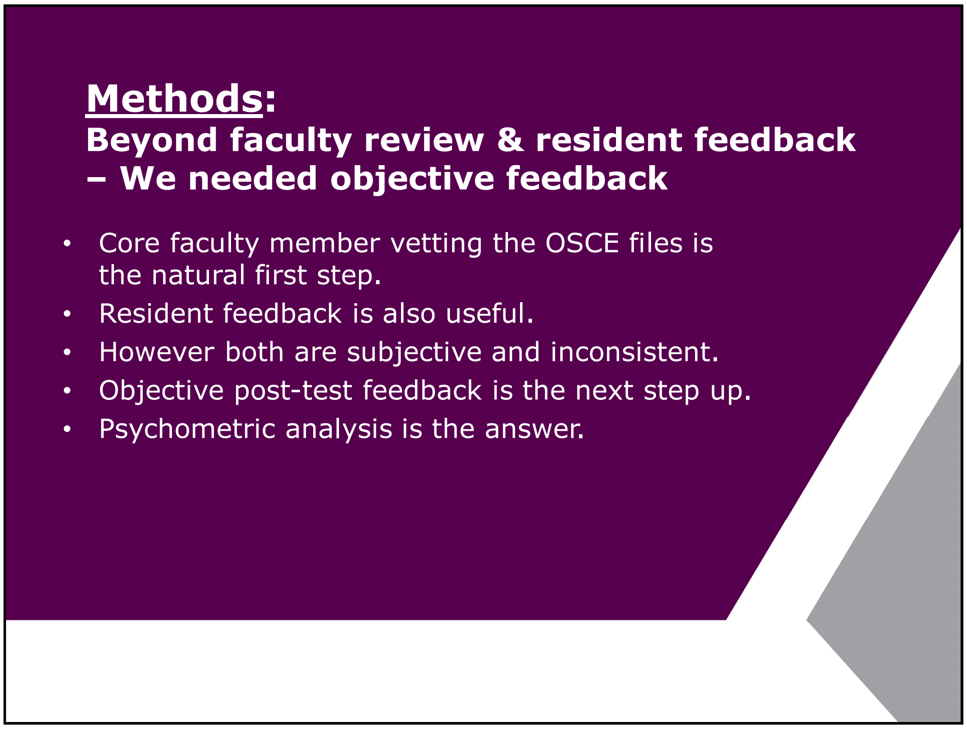
Fig. 1 An example of a film from an OSCE set



Purpose:

Standardise the OSCE sets and improve consistency

- To help the residents with their preparation, the department has a collection of teaching OSCE files.
- These come from different contributors, leading to variations in the difficulty level and quality of these sets.
- This lack of standardisation leads the residents to ‘overcall’ abnormalities not just in the examination, but also in daily clinical practice.



Methods:

Beyond faculty review & resident feedback – We needed objective feedback

- Core faculty member vetting the OSCE files is the natural first step.
- Resident feedback is also useful.
- However both are subjective and inconsistent.
- Objective post-test feedback is the next step up.
- Psychometric analysis is the answer.

Methods:
Psychometric analysis – what is it?

Psychometrics[1]

- A form of quality assurance.
- Provides objective, quantifiable measures.
- Basic psychometric analysis can be performed using simple statistical tests.
- We analysed our OSCEs to ensure that scores were as reliable and valid as possible.

Methods:
Reliability & Validity

Reliability

- Consistency of results.
- Regardless of person/time/situation.
- High reliability across sets implies that the resident would obtain a similar score regardless of which test set he/she took.

Validity

- Is the assessment measuring what is intended?
- High validity of sample sets would imply similar scores on final test.
- High validity of final test would imply similar performance in real life (e.g. emergency dept plain film reporting)

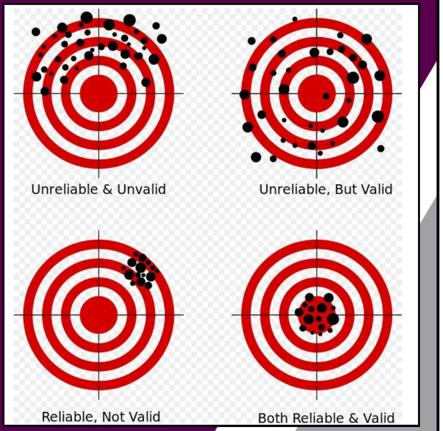
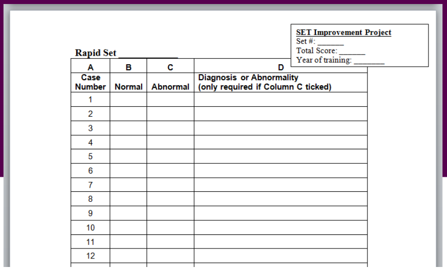


Fig 2. Reliability and validity
© Nevit Dilmen used under under the Creative Commons Attribution-Share Alike 3.0 Unported license.

Methods:
Study methodolgy

- Voluntary participation.
- Anonymized answer sheets collected.
- Following data captured:
  - Year of training, set number
  - Total test score
  - Individual item score
- Analysis with Microsoft Excel 2007 (Redmond, WA)



SET Improvement Project
Set #: _______
Total Score: _______
Year of training: ________

Rapid Set ____________

A	B	C	D
Case Number	Normal	Abnormal	Diagnosis or Abnormality (only required if Column C ticked)
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			

Fig 3. Blank answer form

Methods:
Metrics measured

Test Reliability

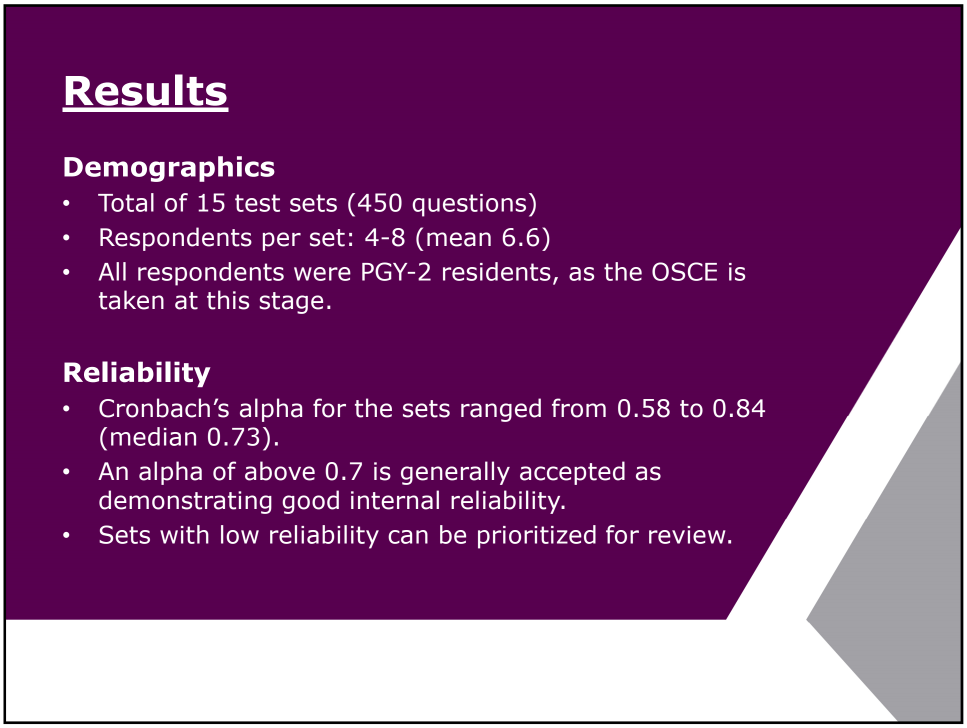
- We chose **Cronbach's Alpha** which has the following advantages:
- Simple and fast to calculate
- Does not need an absolute reference standard e.g. external exam.

Item level metrics

1) Item facility: Percentage of candidates getting that item correct.

2) Item discrimination

- Correlation between performance on an individual question against performance on the overall examination.
- An item with good discrimination would separate the top performing candidates from the poorly performing ones.



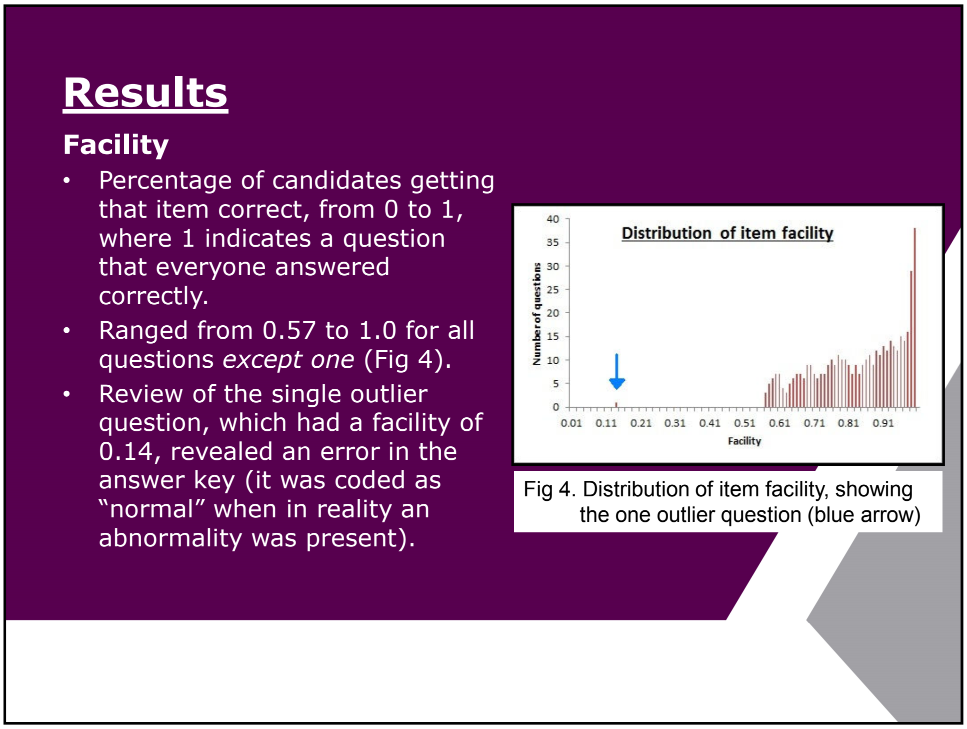
Results

Demographics

- Total of 15 test sets (450 questions)
- Respondents per set: 4-8 (mean 6.6)
- All respondents were PGY-2 residents, as the OSCE is taken at this stage.

Reliability

- Cronbach’s alpha for the sets ranged from 0.58 to 0.84 (median 0.73).
- An alpha of above 0.7 is generally accepted as demonstrating good internal reliability.
- Sets with low reliability can be prioritized for review.



Results

Facility

- Percentage of candidates getting that item correct, from 0 to 1, where 1 indicates a question that everyone answered correctly.
- Ranged from 0.57 to 1.0 for all questions *except one* (Fig 4).
- Review of the single outlier question, which had a facility of 0.14, revealed an error in the answer key (it was coded as “normal” when in reality an abnormality was present).

Fig 4. Distribution of item facility, showing the one outlier question (blue arrow)

Results

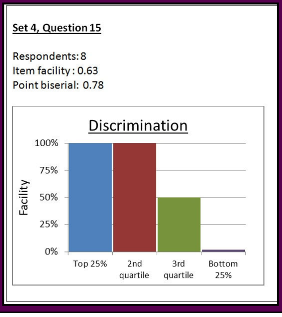
Item Discrimination

- Comparison between performance on an individual question versus performance on the overall test
- Higher values indicate a question that is better able to discriminate between high- and low-performing candidates
- Measured using point biserial correlation coefficients (PBS), which can range from -1 to +1.
- PBS in our series ranged from -0.02 to 0.63. We used low (<0.1) or negative coefficients to identify questions for review.
- Item discrimination can also be expressed visually (Fig 5 and 6).

Results

Item Discrimination

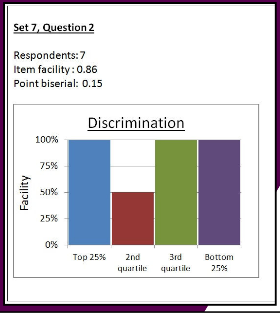
- By dividing the cohort into quartiles, and plotting the facility for each quintile on a bar chart, item discrimination can be represented visually. This works better with larger cohorts.



Set 4, Question 15

Respondents: 8
Item facility: 0.63
Point biserial: 0.78

Fig 5. An item with good discrimination. Note the down-sloping trend from the top 25% of students to the bottom 25%.



Set 7, Question 2

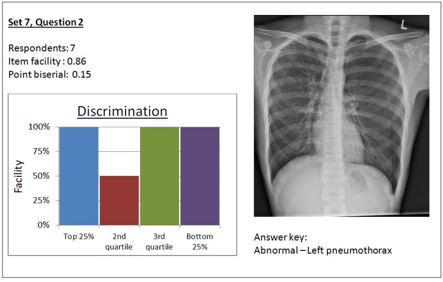
Respondents: 7
Item facility: 0.86
Point biserial: 0.15

Fig 6. An item with poor discrimination, showing lack of correlation between score on the item vs total test score.

Results

Using Item Discrimination + Facility Scores

- We used **low item discrimination scores** to prioritize questions for review, thus allowing us to save manpower and focus our efforts.
- **Facility scores** provided objective evidence for questions that were "too easy" (scores approaching 1.0) or "too hard" (low scores).



Set 7, Question 2

Respondents: 7
Item facility : 0.86
Point biserial: 0.15

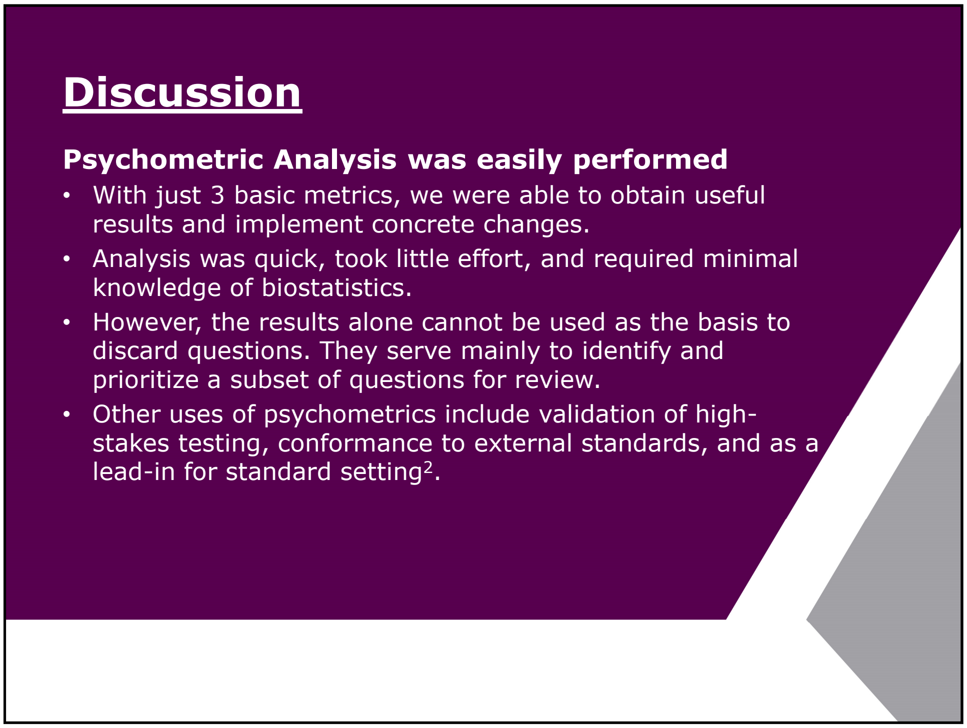
Answer key:
Abnormal – Left pneumothorax

Fig 7. Example of an item with low discrimination and high facility score. Such items were prioritized for review.

Results

Using Item Discrimination + Facility Scores

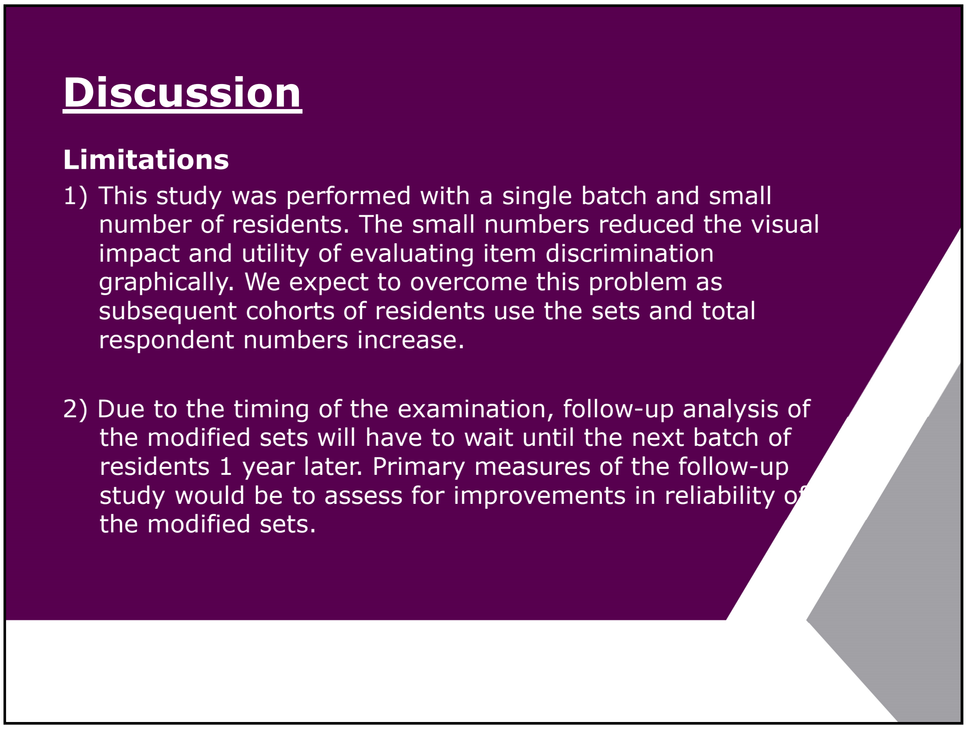
- Note that there is nothing inherently "wrong" with a question having a high or low facility score.
- Instead, our two main objectives in identifying outlier questions were as follows:
- First, we aimed to standardize the difficulty level across the different sets by shifting questions between sets, or by replacing questions (particularly those with low item discrimination scores).
- Second, we were able to identify individual contributors who consistently set "too easy" or "too hard" questions and provide them with objective, evidence-based feedback.



Discussion

Psychometric Analysis was easily performed

- With just 3 basic metrics, we were able to obtain useful results and implement concrete changes.
- Analysis was quick, took little effort, and required minimal knowledge of biostatistics.
- However, the results alone cannot be used as the basis to discard questions. They serve mainly to identify and prioritize a subset of questions for review.
- Other uses of psychometrics include validation of high-stakes testing, conformance to external standards, and as a lead-in for standard setting[2].

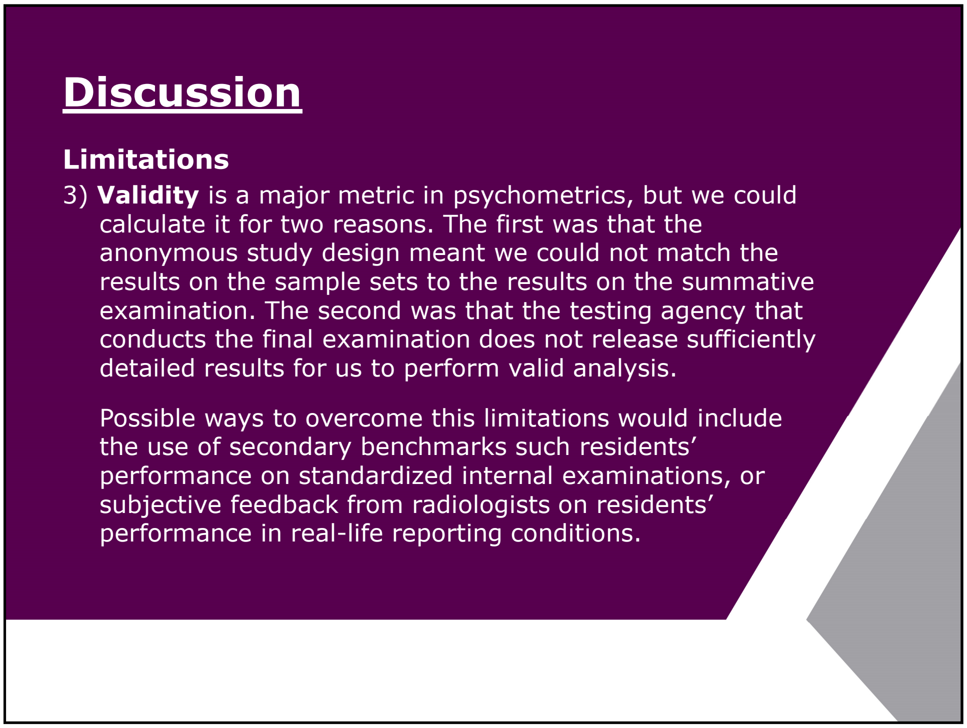


Discussion

Limitations

1) This study was performed with a single batch and small number of residents. The small numbers reduced the visual impact and utility of evaluating item discrimination graphically. We expect to overcome this problem as subsequent cohorts of residents use the sets and total respondent numbers increase.

2) Due to the timing of the examination, follow-up analysis of the modified sets will have to wait until the next batch of residents 1 year later. Primary measures of the follow-up study would be to assess for improvements in reliability of the modified sets.

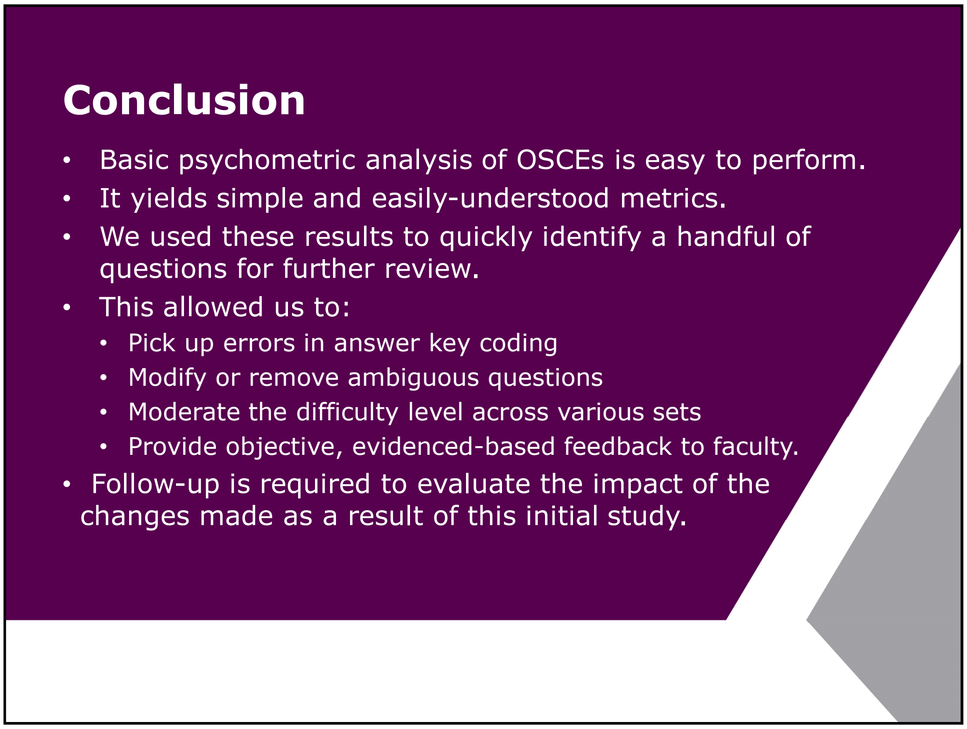


Discussion

Limitations

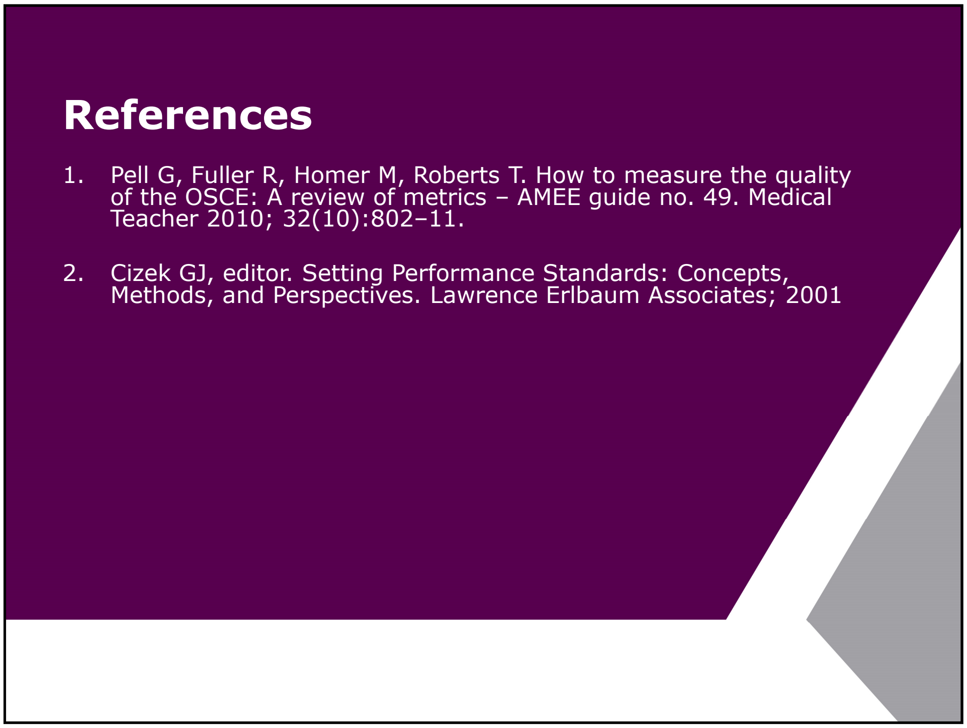
3) **Validity** is a major metric in psychometrics, but we could calculate it for two reasons. The first was that the anonymous study design meant we could not match the results on the sample sets to the results on the summative examination. The second was that the testing agency that conducts the final examination does not release sufficiently detailed results for us to perform valid analysis.

Possible ways to overcome this limitations would include the use of secondary benchmarks such residents' performance on standardized internal examinations, or subjective feedback from radiologists on residents' performance in real-life reporting conditions.



Conclusion

- Basic psychometric analysis of OSCEs is easy to perform.
- It yields simple and easily-understood metrics.
- We used these results to quickly identify a handful of questions for further review.
- This allowed us to:
  - Pick up errors in answer key coding
  - Modify or remove ambiguous questions
  - Moderate the difficulty level across various sets
  - Provide objective, evidenced-based feedback to faculty.
- Follow-up is required to evaluate the impact of the changes made as a result of this initial study.



References

1. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. Medical Teacher 2010; 32(10):802–11.

2. Cizek GJ, editor. Setting Performance Standards: Concepts, Methods, and Perspectives. Lawrence Erlbaum Associates; 2001