Using Performance Exams to Evaluate Curricula

Kevin Volkan, EdD, PhD, MPH
Professor of Psychology
kevin.volkan@csuci.edu

California State University Channel Islands
Camarillo, CA

Presented at the
81st Annual Meeting, April 14, 2005
Accrediting Commission For Senior Colleges & Universities
Western Association of School & Colleges

Using Student Assessment to Evaluate Curricula

Topics

1. Purposes of Assessment
2. Basic Measurement Concepts – Reliability & Validity
3. Different Assessment Methods
4. Performance Assessment
5. OSCE Examples
6. Conclusions

Purposes of Student Assessment

1. Determine Grades – rank students relative to each other, or rank in comparison to an objective criterion
2. Identify students who need help in specific content & performance areas
3. Communicate the importance of the content – *"If its on the test it must be important"*
4. Motivate student learning, skills, and independent learning
5. Diagnose curricular effectiveness & problem areas
6. Identify faculty/teaching problems
7. Identify administrative/organizational problems

Basic Measurement Concepts: Reliability & Validity

Reliability – the stability or consistency of an assessment instrument.

Validity – how well the assessment instrument measures what it is supposed to measure.

Utility – What are the Feasibility, Costs, Practicality, etc.

Basic Measurement Concepts: Reliability & Validity

Reliability

Is greatly affected by the number of items, number of participants, rating scale used, type of reliability coefficient sought, number of raters, time between ratings, type of items, and how well the test questions are written, etc.

Can be affected (usually in a lesser degree) by inconsistencies in the testing location, test conditions, and the presentation of the test itself.

Reliability is really a measure of how much error an assessment carries.

Basic Measurement Concepts: Reliability & Validity

Validity

Is greatly affected by how well the 'domain' being assessed has been defined. This can be examined statistically.

In order for an assessment instrument to be *valid* it must be *reliable*. However, a *reliable* instrument is not necessarily *valid*. Therefore reliability is usually determined first.

There are many ways of determining validity. Usually it is necessary to use multiple methods in determining validity.

Basic Measurement Concepts: Reliability & Validity

An Example of an Assessment that is Reliable but not Valid.

Basic Measurement Concepts: Reliability & Validity

Reliability is usually determined first. It takes many cases/items/raters to generate a reliable instrument.

Validity takes a lot of attention to detail and multiple methods to demonstrate.

A High Quality instrument = an expensive instrument

High Stakes tests need to be very reliable and highly valid. For **Low Stakes** tests this is less important as they are often used more as a learning tool.

Common Rating Problems Related to Reliability/Validity:

Stringency/leniency – Questions or raters can be unnecessarily critical or charitable.

Central tendency - undue avoidance of definitive ratings.

Halo Effect – Basing all ratings on a general impression, or letting one trait overshadow the rest.

Descriptions of Various Assessment Methods

Short-Answer / Essay Questions:

Multiple Choice Questions (MCQs):

Written or Computer Simulations or Cases

Student Portfolios or Log Books

Structured Oral Examinations

Observational Checklists

Performance Exams or OSCE

Electro-Mechanical Simulations

Descriptions of Various Assessment Methods

Short-Answer / Essay Questions:

Description: Describe a problem (usually a case) and probe for understanding/solutions without cueing.

Reliability: Can be OK if there are enough questions and explicit scoring criteria.

Validity: Good if questions are explicit. Difficult to develop sufficiently explicit items that make task clear to examinees.

Utility: Good, though large amounts of examiner and scorer time is needed.

Descriptions of Various Assessment Methods

Multiple Choice Questions (MCQs):

Description: Describe a problem (usually a case) and probe for understanding/solutions without cueing.

Reliability: Excellent. Many situations can be sampled in a short amount of time.

Validity: Can be excellent for assessing cognitive skills, decision-making skills and the ability to apply knowledge to new situations. Using more than five options reduces cueing.

Utility: Costs are high to do correctly, though these are low relative to the benefits.

Descriptions of Various Assessment Methods

Why Using MCQ Tests can be Helpful:

- To reinforce key educational objectives
- To encourage learning
- To assess application of knowledge and skills in analyzing information to solve clinical problems
- To ensure that students read broadly about common and important problems
- To create models of curricular performance
- To gain a ‘snapshot’ of how students learn during their college experience

Descriptions of Various Assessment Methods

Structured Oral Examinations:

Description: Examiners discuss cases/scenarios/problems with students using specified lines of questioning and marking criteria.

Reliability: Often poor because of inadequate case and rater sampling and preparation.

Validity: Poor as it is often unclear what skills are being measured.

Utility: Difficult to do well. Examiners are expensive and the logistics of setting up the exam are resource intensive. Security and equating are problematic. However, these exams are a popular rite of passage at schools around the world.

Descriptions of Various Assessment Methods

Global Rating Scales

Description: Raters (usually faculty provide general indications of quality of knowledge or performance.

Reliability: Internal consistency is often very high, as is test-retest reliability. Inter-rater agreement is often low (but of less importance perhaps).

Validity: Construct validity is often very good as is criterion-related validity after test has been developed. Problems with validity are related to halo effect, rater bias, and limited numbers of raters, or too many raters. Rating can be based on limited observations and/or insufficient quality of observations.

Utility: Costs are very low

Global Rating Scales – *Student Teacher Example*

Please answer the following questions
about each student teacher:

Definitely No Probably No Not Sure Probably Yes Definitely Yes

1. I would take this teacher's suggestions…….

Definitely No Probably No Not Sure Probably Yes Definitely Yes

2. I would place my child with this teacher….

Definitely No Probably No Not Sure Probably Yes Definitely Yes

3. I would recommend this teacher to a friend
……

4. How would you compare the personal manner
(courtesy, respectfulness, sensitivity, friendliness)
of this teacher to other teachers you have seen?

One of the Worst (10%) Below Average (20%) About Average (40%) Above Average (20%) One of the Best (10%)

Descriptions of Various Assessment Methods

“Innovative” Assessment Methods:

- **Performance Exams (eg. Objective Structured Clinical Exams or OSCEs)** allow assessment of basic skills (interviewing, performance, communication, etc.), and ensure exposure to specific problems, including sensitive and difficult problems.
- **Computer Simulations** stimulates interaction among students with faculty, provides corrective feedback and measures ‘real world’ performance.
- **Portfolios/Log Books** document student experiences and can be used to ensure exposure to specific knowledge/problems.

Performance Exams - Objective Structured Clinical Examinations

Description

Trained actors portray the same cases to all examinees. Cases typically focus on history taking, examination, interpretation of data, options/actions to be taken, and communication skills. Raters (usually faculty or actors) evaluate performance (and/or knowledge related to performance) while observing it using observational checklists and/or global rating scales. **Can be applied to many fields which have performance components – Medicine, Nursing, Psychology, Communication, Education, etc.**

Reliability: Depends on rating system used (see global rating scales or observational checklists). Need many cases (9-12 usually is the minimum), many items (usually around 30/case is needed). Also longer exams tend to be more reliable. Raters should either be few or many.

Validity: Usually good for assessing hand-on skills. Potential problems with response-style artifacts, practice effects and rater bias. Experts may also use heuristic shortcuts which the exam may not pick up.

Utility: Expensive and logistically difficult. Security and equating different versions of the exam are difficult. Excellent for providing a learning experience.

Performance Exams – Checklist Example

Station #6: Psychiatry: Student Scoring Checklist

Mark only those items the student performs correctly.

The student will be evaluated on three main areas:

1. The clinical information based on the history, mental status exam, and interview. This should lead to a primary diagnostic impression and a differential diagnostic with appropriate "rule-outs."

2. A brief summary of the appropriate acute intervention(s) – i.e. the important elements of the patient's care that should be addressed today.

3. The student's interactions with the patient.

1. History, Mental Status Exam, and Differential Diagnosis

The student must address the following clinical information in the interview (are these items present or absent in the patient?) in order to make the diagnosis of mood disorder and rule out other conditions:

A. History, past and recent:

Did the student ask about the following items?

Mark if YES:

- ○ Any previous history of depression, hypomania, mania, suicidality, etc.
- ○ Any change in appetite or weight
- ○ Any change in sleep pattern ……etc, etc.

Objective Structured Clinical Examinations

Why were OSCEs developed?

a. It was felt that written tests could not measure performance competence

b. There are practical problems with rating student encounters in the real world - these are subject to a lot of bias – problem variability and availability is a problem, not to mention subjecting public to student performances.

d. Having more advanced students role play as actors in cases has problems

e. Having faculty role play as s actors in cases also has problems and is expensive

Objective Structured Clinical Examinations

What can OSCEs do?

OSCEs can measure clinical competence with good reliability and some validity in the following areas:

- Interview skills
- Knowledge related to ‘hands on’ assessment of situation
- Technique related to ‘hands on’ assessment of situation
- Diagnostic skills
- Ability to reason *in vivo* (think on feet)
- Case management skills
- Communication skills

OSCEs can serve an important function as a teaching tool that allows for immediate feedback on skills

Hamann C, Volkan K, Fishman MB, et al. (2002). How well do second-year students learn physical diagnosis? Results from an Objective structured clinical examination (OSCE).*BMC Medical Education* (2002) 2:1

Objective Structured Clinical Examinations

What can OSCEs do for Faculty Efforts?

OSCEs can be used to diagnose quality and variability in a curriculum with regard to performance skills. More specifically, OSCEs can give an indication of:

- How well a course is teaching a particular skill that students must perform

- Consistency of faculty teaching efforts

- How consistent performance skills are being taught across different sites

Morag E, Lieberman G, Volkan K, et al (2001). Clinical competence assessment in radiology. *Academic Radiology*, *8*, 74-81

What can OSCEs Tell You About Student Outcomes?

OSCEs can be used to diagnose quality and variability among students. OSCEs can also give an indication of the quality/relevance of other student assessment efforts:

- Future performance of an individual student or a class cohort
- Whether student outcomes are being met
- The influence of a specific course or skill on subsequent student performance
- The relationship of student characteristics and skill sets to success in the curriculum
- The relationship of success in the curriculum to external (state licensing, professional group) assessments of performance/knowledge

Simon, SR., Volkan, K., Hamann, C., Duffey, C., & Fletcher, SW. (2002). The relationship between second-year medical student's OSCE scores and the USMLE Step 1 scores. Medical Teacher, 24(5), pp. 535-539

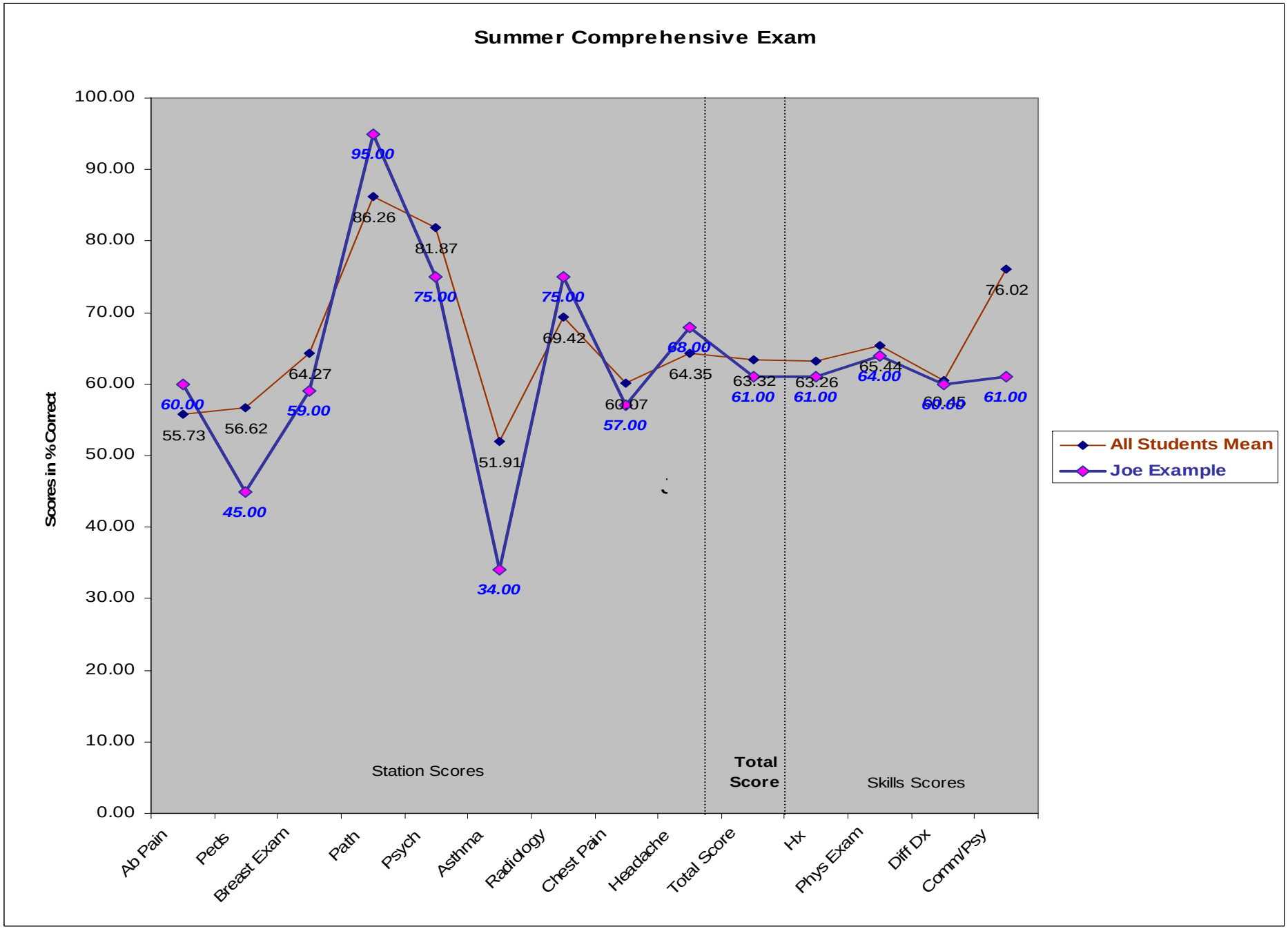
How can OSCEs Enhance Student Learning?

- Can give individual student or a class cohort a sense of how they performed compare to rest of class, or other sample

- Can indicate specific areas needing improvement

- Gives the student a chance to get intensive detailed and personal feedback from faculty – often senior faculty!

Film Clip – back pain case


Summer Comprehensive Exam
Scores in % Correct
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
55.73
56.62
64.27
86.26
81.87
51.91
69.42
60.07
64.35
63.32
63.26
65.44
60.45
76.02
60.00
45.00
59.00
95.00
75.00
34.00
75.00
57.00
68.00
61.00
61.00
64.00
60.00
61.00
Station Scores
Total Score
Skills Scores
Ab Pain
Peds
Breast Exam
Path
Psych
Asthma
Radiology
Chest Pain
Headache
Total Score
Hx
Phys Exam
Diff Dx
Comm/Psy
All Students Mean
Joe Example

The last two can sometimes get you into trouble!

The relationship of student characteristics and skill sets to success in the curriculum

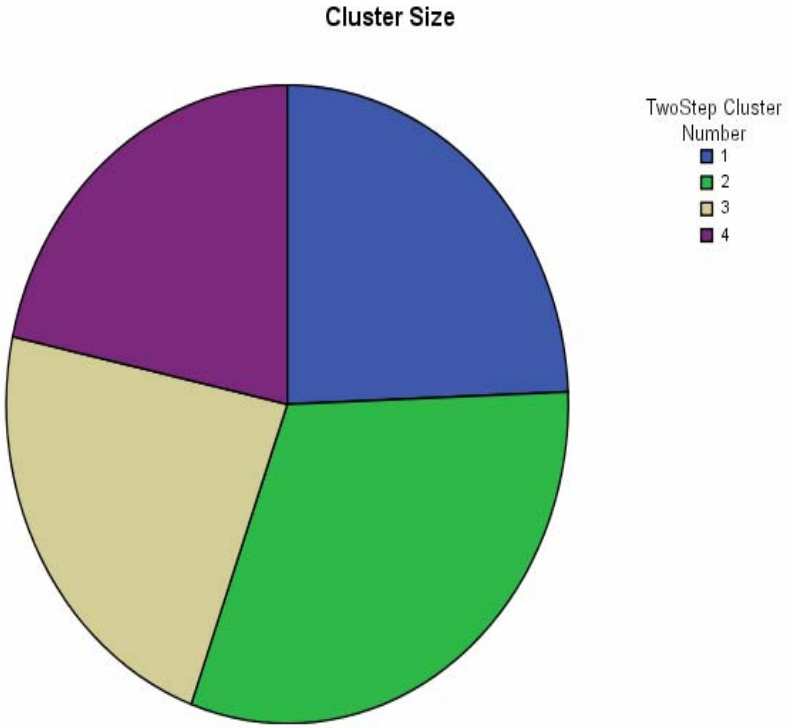
The relationship of success in the curriculum to external (state licensing, professional group) assessments of performance/knowledge

Gender numeric

		.00		1.00	
		Frequency	Percent	Frequency	Percent
Cluster	1	0	.0%	79	43.9%
	2	0	.0%	101	56.1%
	3	74	51.4%	0	.0%
	4	70	48.6%	0	.0%
	Combined	144	100.0%	180	100.0%

White vs Other

		Non-White		White	
		Frequency	Percent	Frequency	Percent
Cluster	1	79	53.0%	0	.0%
	2	0	.0%	101	57.7%
	3	0	.0%	74	42.3%
	4	70	47.0%	0	.0%
	Combined	149	100.0%	175	100.0%

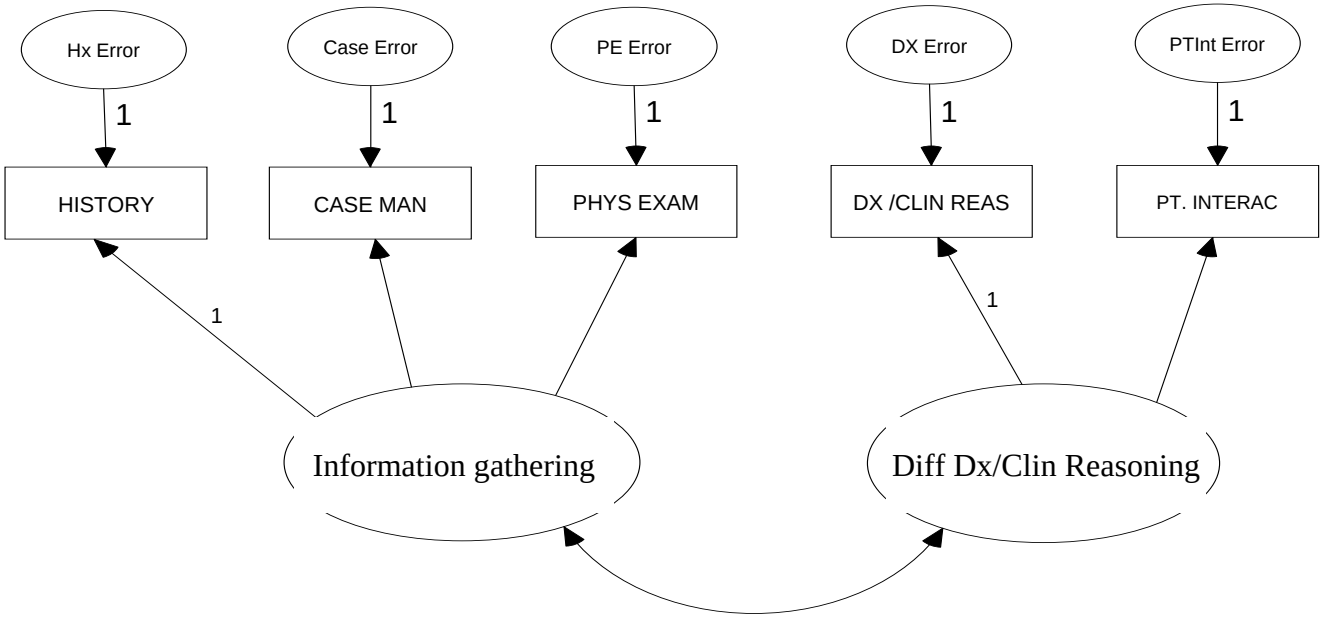


Centroids

		step2		NBME Step 1 Score		Total Score	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
cluster	1. NW- Female	224.99	21.951	229.6456	18.59559	57.7975	5.98370
	2. W-Female	236.94	16.965	237.1386	14.52035	59.7525	4.65275
	3. W-Male	229.26	22.673	237.6757	16.17042	59.1757	5.51053
	4. NW-Male	222.53	23.141	232.5286	17.46526	57.6000	5.58570
	Combined	229.16	21.651	234.4383	16.85614	58.6790	5.45304

Objective Structured Clinical Examinations

May be able to measure both *information gathering* and *deductive reasoning*



Two Factor Structural Model of Comprehensive Exam OSCE

Volkan, K., Simon, S., Baker, H., & Todres, I.D. (2004). Psychometric Structure of a Comprehensive Objective Structured Clinical Examination: A Factor Analytic Approach. *Advances in Health Sciences Education, 9* (2): 83-92

The ability to measure both information gathering and deductive reasoning may be key to providing students with useful feedback about their performance.

It also shows the relationship between performance and knowledge leading to more precise and helpful feedback to students.

Challenges of Using Performance Assessments

Challenges

Expensive!

Difficult to do right – like producing a movie!

Tough to get faculty to contribute time

Potential volume of data can be overwhelming

Statistical challenges to analyzing and understanding the data – multiple sources of error

Important Points

1. Performance assessments measure the integrated application of skills and knowledge.
2. Formative feedback from these methods can be used to enhance student learning and to revise academic programs.
3. Student performance can be predictive of future professional success.
4. Measurement of knowledge is not necessarily measurement of performance or skills. BUT measurement of performance often measures some knowledge.