VALIDITY AND RELIABILTY

Compiled Resources from an online discussion carried out by CMCL-FAIMER fellows of the Class of 2007 in March 2007

Moderators : Dr Anshu and Dr Chetna Desai
Faculty : Dr Payal Bansal and Dr Tejinder Singh

CONTENTS

SESSION ONE:

INTRODUCTION

✉ **Compiled by Anshu:**
We are glad to kickstart the beginning of the online discussions on pertinent issues in medical education. The first theme we had chosen was to unravel the confusion that exists about two terms in assessment- validity and reliability.

We have chosen to spend the next four weeks deliberating this issue under the following heads:

Week One	: Introduction to terms and use in assessment
Week Two	: Deficiencies in our present systems of evaluation Factors which affect validity and reliability
Week Three	: Strengths and weaknesses of assessment tools
Week Four	: Improving validity and reliability of assessment tools Improving assessment in curriculum innovation projects

INTRODUCTION TO TERMS

Measurement refers to application of mathematical tools for finding degree of achievement. E.g. MCQs

Assessment is used for those attributes which cannot be precisely measured and where some degree of subjectivity is involved. E.g. Essay-type questions

Evaluation is the process of determining whether pre-determined educational objectives have been achieved. It involves passing a value judgement based on the information obtained by measurement and assessment.

- **Formative or diagnostic evaluation** is used to serve as a feedback to learners and teachers. This feedback can be used to correct learning methods and teaching styles. It should not count for pass or fail decisions.

- **Internal evaluation** is a term used when evaluation is carried out by the teacher who has taught the subject himself. This has to be continuous to be meaningful.

- **Summative evaluation** is end of course evaluation to know if the student is competent enough for certification. Ideally summative evaluation should be

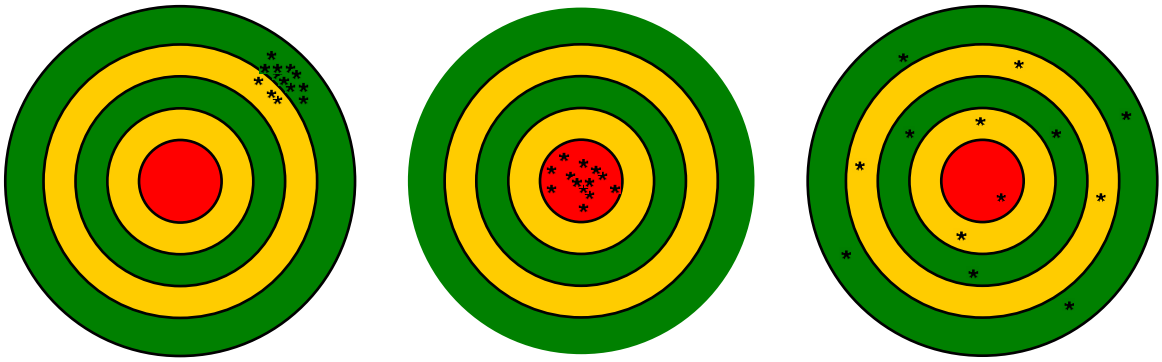
criterion based. However on most occasions it is usually norm-referenced i.e. students are evaluated by comparison with their peers.

Characteristics of evaluation tools

Relevance	Is it appropriate in the context of needs?
Validity	Does it test what it is intended to test?
Reliability	Does it consistently test what is intended to test?
Objectivity	Will scores obtained by candidate be same if examined by two independent expert examiners?
Feasibility	Can it be implemented in practice?

VALIDITY AND RELIABILITY

- Validity refers to the meaning of the test scores.
- Reliability refers to the consistency of the measurement. It is a measure of reproducibility of a test.



Consistent but not accurate ⇩ Reliable but not valid

Consistent and accurate ⇩ Reliable and valid

Neither consistent nor accurate ⇩ Neither reliable nor valid

VALIDITY:

Does the instrument measure the property that it intends to measure?

Validity is the most important quality to consider in assessment and is concerned with

- the appropriateness,
- meaningfulness and
- usefulness

of the specific inferences made from the assessment results.

Validity is concerned with the interpretation and use of assessment results.

E.g.: If you intend to determine if students know how to elicit the knee jerk, you can
(a) Ask them to write down the steps of eliciting a knee jerk OR
(b) Ask them to demonstrate on a patient.

Both are valid methods, but (b) is a more valid method than (a).

Types of validity

In the past, validity was defined as three different types: content, criterion and construct. Now construct validity is considered the sole type of validity

1. CONTENT VALIDITY:

Does the test measure what was intended to be taught?
Content validity is ensured by the following steps:

- Prepare list of content matter or behavioural changes to be tested
- Assign weightage to above based on relative importance
- Prepare a table of specifications
- Create test according to table of specifications

2. CRITERION VALIDITY:

Do persons who are taught do better than those not taught?
It refers to validity in relation to an external criterion. Unlike content validity, criterion validity can be numerically expressed.
It is a comparison of test scores on the same behaviour

- obtained by some other measuring instrument at the same time **(Concurrent validity)** or
- with subsequent scores obtained by another test given at some future time **(Predictive validity)**

3. CONSTRUCT VALIDITY:

Do scores in the test correlate with presence of other qualities expected to be related to that which is being tested?

Construct validity is really a comparison of the test scores with scores obtained by other measuring instruments which measure a related ability which is usually associated with that tested by the current measuring instrument.
E.g.: A test which seeks to measure a student's history taking skill maybe expected to correlate with scores of a test which measures his interviewing skills in general.

Constructs refer to collections of abstract concepts and principles which are inferred from behaviour and explained by educational and psychological theory. Nearly all assessments in medical education deal with constructs.

Educational achievement is a construct- which is inferred from performance in written tests, oral examinations and standardized patient examinations. Educational ability or aptitude is a construct which is more abstract than achievement because there is less agreement about its meaning among educators.

If we infer that the assessment is useful for predicting or estimating some other performance, we would like some credible evidence to support that interpretation.

Example:
Use of validity and reliability in entrance selection exams: While selecting prospective students for admission in medical schools, one must have scientifically sound evidence that the entrance exam scores reflect the selection criteria one is looking for. One way is to compare entrance exam scores with achievement in medical school at a later date.
Thus we require a chain of evidence with theory, hypotheses and logic to support or refute the reasonableness of interpretations. High pass/fail decision reliability is essential for high stakes examinations.

Validity has the following characteristics.

1. Validity is never assumed. Validity is *inferred* from available evidence(not measured)
2. Validity requires multiple sources of evidence to support or refute meaningful score interpretation
3. Validity is expressed by *degree.* Assessment is never valid or invalid; rather scores have more (or less) validity.
4. Validity is specific to a particular *use*
5. Validity refers to the *inferences drawn*, not the instrument.
6. Validity is a *unitary concept*
7. Validity is concerned with the *consequences of using the assessments*

RELIABILITY

- Extent to which test produces the same results when used repeatedly under the same conditions
- Consistency of a measurement made with a particular test
- Consistency of scores upon repeated measurement of the same individuals

Reliability refers to the results of a test and not to the instrument itself. A reliable test is not necessarily a valid one. A test can consistently measure something, but not necessarily what is intended to measure.

Reliability is therefore, necessary, but not by itself sufficient to ensure validity.

Unlike validity (where some subjective judgement maybe involved), reliability is a strictly mathematical concept and is a measure of correlation between two sets of scores

Types of Reliability

1. TEST-RETEST RELIABILITY:

This is to administer the same test after an interval and compare the scores- provided no additional learning has occurred and nothing has been forgotten in that interval. This is almost impossible to test in practice. Repeatedly giving the same test may produce distorted results.

2. EQUIVALENT TESTS:

Compare two tests of equivalent form (same content, same difficulty level) which is administered to the same group of students to obtain two sets of scores

3. SPLIT HALF METHOD

A single test is split into two halves and the two sets of scores for each student are compared. It is a measure of internal consistency of the test.

4. MARKER RELIABILITY:

The degree of consistency when a test paper is independently marked by two different examiners

DISCUSSION TOPICS FOR WEEK ONE:

We now invite your opinions and views on the following issues:

1. What do you think are the most appropriate assessment tools in different learning areas of your subject? Please discuss evaluation of cognitive, psychomotor and affective domains with examples.

2. Which according to you are the most appropriate assessment tools in the following situations:
   (a) Pre-testing of students before beginning a series of lectures to assess their basic knowledge
   (b) Assessment of ability of a student to pursue self learning after passing final MBBS
   (c) Ability of a student to defend his clinical diagnosis
   (d) Ability to work as an effective member of a health team
   (e) Ability to reassure parents of a child who has a bad prognosis
   (f) Ability of a student to carry out health education activities in the community
   (g) Ability of a student to carry out a psychomotor skill.

3. What are the appropriate assessment tools in your curriculum innovation project?

COMPILED RESPONSES FROM FELLOWS AND FACULTY

The concept of reliability and validity

✉ **Bill Burdick:** Our ideal is a test result that is the same with repeated measures, and that means what we say it means.

The same test result with repeated measures? Let's look at this more closely -

If I could wipe out my memory and take the same test again tomorrow, I should get the same result. Why wouldn't I get the same result? Some factors would be related to me (the examinee) – I might have a slight headache today, I might have had an argument with my girlfriend last night distracting me. Other factors might be related to the environment – there might be construction today outside the building making a distracting noise, maybe the AC is set too high today. Other factors might relate to the test itself – there might be a different examiner scoring the essay question, or giving the oral exam, or they might chose a different patient, or the standardized patient in the OSCE might act a little differently. This is the concept of reliability.

The test results mean what we say they mean? Let's look at this more closely –

My teacher gives me a written test at the end of my statistics course. The test asks for definitions of statistical terms. My teacher concludes from these results that I have familiarity with statistical terminology. This inference (conclusion) sounds valid, and if we wanted evidence for the validity of this inference we might compare this test result with another assessment of my familiarity with statistical terms – perhaps ask faculty who know me, or perhaps ask me to orally explain the terms. With this evidence, I could present a credible argument for the validity of this inference. However, if my teacher used the same test results and from them concluded that I am able to calculate confidence intervals for a lab result, I would be very skeptical of the validity of that inference. I could even develop evidence to refute this claim. This is the concept of validity.

✉ **Stewart Mennin**: I have to remind myself that validity is about inferences we make from information and reliability is about consistency of measurement of information.

The reliability of a test describes the degree to which the test consistently measures what it is supposed to measure. The more reliable a test, the more likely it is that a similar result will be obtained if the test is re-administered. Reliability is sensitive to the length of the test, the station or item discrimination, and the heterogeneity of the cohort of candidates. Standardized patients' portrayals, patients' behaviour, examiners' behaviour, and administrative variables also affect reliability.

✉ **Venugopal Rao:**

From http://writing.colostate.edu/guides/research/relval/index.cfm

Validity

- The extent to which an assessment procedure measures what its authors or users claim it measures
- According to Joint Committee on Test Standards, the appropriateness, meaningfulness and usefulness of the specific inferences made from a test score
- Validity is a property of test-based inferences and not a property of the test itself.
- Test validation is process of accumulating evidence to support such inferences
- Regardless of how evidence collected, validity always refers to the degree to which the evidence supports the inferences that are made from the scores.

Validity refers to the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure. While reliability is concerned with the accuracy of the actual measuring instrument or procedure, validity is concerned with the study's success at measuring what the researchers set out to measure.

Researchers should be concerned with both *external* and *internal* validity. External validity refers to the extent to which the results of a study are generalizable or transferable.

Internal validity refers to (1) the rigor with which the study was conducted (e.g., the study's design, the care taken to conduct measurements, and decisions concerning what was and wasn't measured) and (2) the extent to which the designers of a study have taken into account alternative explanations for any causal relationships they explore (Huitt, 1998). In studies that do not explore causal relationships, only the first of these definitions should be considered when assessing internal validity.

Scholars discuss several types of internal validity.

- Face Validity
- Criterion Related Validity
- Construct Validity
- Content Validity

Validity: Example

Many recreational activities of high school students involve driving cars. A researcher, wanting to measure whether recreational activities have a negative effect on grade point average in high school students, might conduct a survey asking how many students drive to school and then attempt to find a correlation between these two factors. Because many students might use their cars for purposes other than or in addition to recreation (e.g., driving to work after school, driving to school rather than walking or taking a bus), this research study might prove invalid. Even if a strong correlation was found between driving and grade point average, driving to school in and of itself would seem to be an invalid measure of recreational activity.

Face Validity

Face validity is concerned with how a measure or procedure appears. Does it seem like a reasonable way to gain the information the researchers are attempting to obtain? Does it seem well designed? Does it seem as though it will work reliably? Unlike content validity, face validity does not depend on established theories for support

Criterion Related Validity

Criterion related validity, also referred to as instrumental validity, is used to demonstrate the accuracy of a measure or procedure by comparing it with another measure or procedure which has been demonstrated to be valid.

For example, imagine a hands-on driving test has been shown to be an accurate test of driving skills. By comparing the scores on the written driving test with the scores from the hands-on driving test, the written test can be validated by using a criterion related strategy in which the hands-on driving test is compared to the written test.

Construct Validity

Construct validity seeks agreement between a theoretical concept and a specific measuring device or procedure. For example, a researcher inventing a new IQ test might spend a great deal of time attempting to "define" intelligence in order to reach an acceptable level of construct validity.

Construct validity can be broken down into two sub-categories: Convergent validity and discriminate validity. Convergent validity is the actual general agreement among ratings, gathered independently of one another, where measures should be theoretically related. Discriminate validity is the lack of a relationship among measures which theoretically should not be related.

To understand whether a piece of research has construct validity, three steps should be followed. First, the theoretical relationships must be specified. Second, the empirical relationships between the measures of the concepts must be examined. Third, the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure being tested (Carmines & Zeller, p. 23).

Content Validity

Content Validity is based on the extent to which a measurement reflects the specific intended domain of content (Carmines & Zeller, 1991, p.20).

Content validity is illustrated using the following examples: Researchers aim to study mathematical learning and create a survey to test for mathematical skill. If these researchers only tested for multiplication and then drew conclusions from that survey, their study would not show content validity because it excludes other mathematical functions. Although the establishment of content validity for placement-type exams seems relatively straight-forward, the process becomes more complex as it moves into the more abstract domain of socio-cultural studies. For example, a researcher needing to measure an attitude like self-esteem must decide what constitutes a relevant domain of content for that attitude. For socio-cultural studies, content validity forces the researchers to define the very domains they are attempting to study.

Clarifying the Concept of Construct Validity

✉ **Hemlata Badyal:** Please clarify construct validity with an example.

✉ **Payal Bansal:** It suffices to understand that the term "Construct" in our context is *educational achievement:*
For example, the Handout on X-Ray interpretation, X-Ray interpretation is the construct.

We can use various tools to measure this construct, e.g. giving one X-Ray, giving many X-Rays, X-Rays followed by oral viva, X-Ray followed by written questions (simple or complex), X-Rays of one system, or of many system.

When we look for "validity evidence", we look for evidence to support that the student's score truly reflects that ***that*** educational achievement (Ability to interpret X-Rays) has taken place. E.g. if a student scores a certain score (say, 7 or 8 (or whatever standard we decide) out of 10), we can say confidently that he can interpret X-Rays appropriate to his level. The more "evidence of validity" that we have, the more confidently we can say this.

Please refer to the **handout** for what kind of evidence represents what kind of validity.

Other constructs can be the different competencies that we expect from the student, for example, level of knowledge and understanding at the beginning of a course.

The entry level MCQ exam for entrance tests is a way of measuring this construct so, what is the validity evidence that we will look for when we say that this MCQ truly reflects what competence we are looking for at the entry level. Is the MCQ test enough or should other tools be combined? Is the quality of the test upto our satisfaction (number of questions, type of questions, range of questions)? Was difficulty level appropriate? There are other questions to be asked, each reflecting a specific type of validity evidence that contributes to the validity as a whole. The attachment should be a good guide.

Other examples of construct or educational achievement-
Ability to take a complete history
Ability to interpret lab data
Ability to interpret/identify a given set of microscopic slides
Ability to explain pathophysiology

Reliability

✉Chandrika Rao:

- A test is not reliable, a score is. Reliability is a characteristic of data, not the instrument
- Refers to the stability (reproducibility and consistency) of a measure
- In classical theory reliability is the amount of error present in the scores yielded by a test.
- According to Joint Committee on test Standards, “reliability refers to the accuracy (consistency and stability) of measurement by a test.”
- Reliability is a measure of the extent to which random variation affects the measurement of a trait, characteristic, or quality
- Reliability is degree to which test scores are free from errors of measurement (APA, AERA, & NCME, 1985)

✉Venugopal Rao:

For researchers, four key types of reliability are:

- Equivalency Reliability
- Stability Reliability
- Internal Consistency
- Interrater Reliability

Equivalency Reliability

Equivalency reliability is the extent to which two items measure identical concepts at an identical level of difficulty. Equivalency reliability is determined by relating two sets of test scores to one another to highlight the degree of relationship or association. In quantitative studies and particularly in experimental studies, a correlation coefficient, statistically referred to as *r*, is used to show the strength of the correlation between a dependent variable (the subject under study), and one or more independent variables, which are manipulated to determine effects on the dependent variable. An important consideration is that equivalency reliability is concerned with correlational, not causal, relationships.

For example, a researcher studying university English students happened to notice that when some students were studying for finals, their holiday shopping began. Intrigued by this, the researcher attempted to observe how often, or to what degree, these two behaviors co-occurred throughout the academic year. The researcher used the results of the observations to assess the correlation between studying throughout the academic year and shopping for gifts. The researcher concluded there was poor equivalency reliability between the two actions. In other words, studying was not a reliable predictor of shopping for gifts.

Stability Reliability

Stability reliability (sometimes called test, re-test reliability) is the agreement of measuring instruments over time. To determine stability, a measure or test is repeated on the same subjects at a future date. Results are compared and correlated with the initial test to give a measure of stability.

An example of stability reliability would be the method of maintaining weights used by the U.S. Bureau of Standards. Platinum objects of fixed weight (one kilogram, one pound, etc...) are kept locked away. Once a year they are taken out and weighed, allowing scales to be reset so they are "weighing" accurately. Keeping track of how much the scales are off from year to year establishes stability reliability for these instruments. In this instance, the platinum weights themselves are assumed to have perfectly fixed stability reliability.

Internal Consistency

Internal consistency is the extent to which tests or procedures assess the same characteristic, skill or quality. It is a measure of the precision between the observers or of the measuring instruments used in a study. This type of reliability often helps researchers interpret data and predict the value of scores and the limits of the relationship among variables.

For example, a researcher designs a questionnaire to find out about college students' dissatisfaction with a particular textbook. Analyzing the internal consistency of the survey items dealing with dissatisfaction will reveal the extent to which items on the questionnaire focus on the notion of dissatisfaction.

Interrater Reliability

Interrater reliability is the extent to which two or more individuals (coders or raters) agree. Interrater reliability addresses the consistency of the implementation of a rating system.

A test of interrater reliability would be the following scenario: Two or more researchers are observing a high school classroom. The class is discussing a movie that they have just viewed as a group. The researchers have a sliding rating scale (1 being most positive, 5 being most negative) with which they are rating the student's oral responses. Interrater reliability assesses the consistency of how the rating system is implemented. For example, if one researcher gives a "1" to a student response, while another researcher gives a "5," obviously the interrater reliability would be inconsistent. Interrater reliability is dependent upon the ability of two or more individuals to be consistent. Training, education and monitoring skills can enhance interrater reliability.

Difficulties of Achieving Reliability

It is important to understand some of the problems concerning reliability which might arise. It would be ideal to reliably measure, every time, exactly those things which we intend to measure. However, researchers can go to great lengths and make every attempt to ensure accuracy in their studies, and still deal with the inherent difficulties of measuring particular events or behaviors. Sometimes, and particularly in studies of natural settings, the only measuring device available is the researcher's own observations of human interaction or human reaction to varying stimuli. As these methods are ultimately subjective in nature, results may be unreliable and multiple interpretations are possible. Three of these inherent difficulties are quixotic reliability, diachronic reliability and synchronic reliability.

Quixotic reliability refers to the situation where a single manner of observation consistently, yet erroneously, yields the same result. It is often a problem when research appears to be going well. This consistency might seem to suggest that the experiment was demonstrating perfect stability reliability. This, however, would not be the case.

For example, if a measuring device used in an Olympic competition always read 100 meters for every discus throw, this would be an example of an instrument consistently, yet erroneously, yielding the same result. However, quixotic reliability is often more subtle in its occurrences than this. For example, suppose a group of German researchers doing an ethnographic study of American attitudes ask questions and record responses. Parts of their study might produce responses which seem reliable, yet turn out to measure felicitous verbal embellishments required for "correct" social behavior. Asking Americans, "How are you?" for example, would in most cases, elicit the token, "Fine, thanks." However, this response would not accurately represent the mental or physical state of the respondents.

(✉Anshu: One 'reliable' professor I know, gives 6 marks out of 10 to all 60 students in their assessment! My jaw drops at his zeal to be consistent and fair to his students, and pass all of them.

One measuring instrument which I really wish would show similar consistency- are my bathroom scales. I'm tired of their swinging precariously to the right everytime I step on them. Couldn't they reliably be stuck on one low score for heaven's sake?! **)**

Diachronic reliability refers to the stability of observations over time. It is similar to stability reliability in that it deals with time. While this type of reliability is appropriate to assess features that remain relatively unchanged over time, such as landscape benchmarks or buildings, the same level of reliability is more difficult to achieve with socio-cultural phenomena.

For example, in a follow-up study one year later of reading comprehension in a specific group of school children, diachronic reliability would be hard to achieve. If the test were given to the same subjects a year later, many confounding variables would have impacted the researchers' ability to reproduce the same circumstances present at the first test. The final results would almost assuredly not reflect the degree of stability sought by the researchers.

Synchronic reliability refers to the similarity of observations within the same time frame; it is not about the similarity of things observed. Synchronic reliability, unlike diachronic reliability, rarely involves observations of identical things. Rather, it concerns itself with particularities of interest to the research.

For example, a researcher studies the actions of a duck's wing in flight and the actions of a hummingbird's wing in flight. Despite the fact that the researcher is studying two distinctly different kinds of wings, the action of the wings and the phenomenon produced is the same.

***

Which is more important: Validity or reliability?

✉ **Dinesh Badyal:** I think both are equally important. Think of a scenario when a method is valid but not reliable, or when a method is reliable but not valid. A little bit of compromise may be there, but both should have equal importance. Weightage given may be different for both.

✉ **Tejinder Singh:** Let me give an example. PG entrance tests based on MCQ are reliable, but are they valid? Are you comfortable selecting PGs on the basis of MCQs only? Any comments now?

✉ **Chandrika Rao:** These two properties, validity and reliability, are discussed in nearly all texts on instructional testing. They are usually mentioned in the opposite order--reliability and validity. However, even though there are good reasons for that order, it is a mistake because it underemphasizes validity, which is, by far, the more important of the two. It is also the harder to gather evidence about.

✉ **Vivek Saoji:** Validity is more important as it means the meaningful interpretation of the test data and that can we meaningfully infer from the test scores. In validity we gather evidence from multiple sources either in favour of or against validity. Reliability is one of the validity evidences and therefore a part of validity.

Test scores may be reliable but may not be valid. But if they are valid they have to be reliable. Therefore though both are important, validity is more important and many purists (theory) believe that there is no need to discuss reliability separately as it is a part of validity.

✉ **Payal Bansal:** If reliability is more, it contributes more to validity evidence. If less, it contributes to a lesser degree. If reliability is really low, the defence for validity becomes really weak. Reliability should be looked upon as a part of validity.

✉ **Tejinder Singh:** There are situations, when you may have to choose between one and in that case, always go for validity. Most of the times, we go for reliability, ignoring validity. A valid test is generally reliable (not always) but a reliable test may not always be valid.

E.g. Using shoe size to measure intelligence: It is a highly reliable measure as everytime, the size will be the same but can you use it to predict intelligence? No.
Using IQ test for measuring intelligence may not be as reliable, because the results may vary depending on the tester, mood of the student, time of the day and so on- but you would rather use this than shoe size.

That is the point, which often is forgotten. As you go higher in learning domains, (decision making, interpersonal skills, communication for example), the evaluation tools become less reliable. Is it fair to settle for a more reliable tool but less valid tool?

✉ **Payal Bansal:** If reliability is poor, the support for validity is weak, and so it will not be appropriate to use the scores for making a major decision about the candidate. Better to look for the best available tool likely to yield reliable scores.

✉**Dinesh Badyal**: If the situation demands, we can go for a tool which is less reliable, but is valid.

✉ **Tejinder Singh:** It is unfair to the student to take any decision based on unreliable results. Having said that, the solution lies in making the results more reliable. (Remember the premise- it is the results and not tests, which are reliable or unreliable. If for example, you find that the results of direct observation of clinical performance are not reliable, then make the results more reliable by increasing the number of observers, number of observations, breaking the observed task into components, structuring the observations, using check lists and so on but as said previously, you can not use shoe size for clinical performance, howsoever reliable it may be. If you are concerned about low reliability, improve the quality of the tool to make results more reliable.

Can a perfectly valid test be unreliable?

✉ **Chetna Desai:**

- Are there any examples where a perfectly valid test may not be reliable?
- In such cases are we allowed to adopt these tests for research?

✉ **Payal Bansal:** The answer is no, because there is no "perfectly valid test" or "perfectly valid test scores". But for a score to be a valid interpretation as a pass/fail decision, some level of reliability is required. So either use a better tool or improve the tool to increase its reliability. Both options are there.

Measuring Validity and Reliability

✉ **Chandrika Rao:** Two of the primary criteria of evaluation in any measurement or observation are:

1. Are we are measuring what we intend to measure?
2. Does the same measurement process yield the same results?

These two concepts are validity and reliability.

Reliability is concerned with questions of stability and consistency - does the same measurement tool yield stable and consistent results when repeated over time. Think about measurement processes in other contexts - in construction or woodworking, a tape measure is a highly reliable measuring instrument.

Say you have a piece of wood that is 2.5 feet long. You measure it once with the tape measure - you get a measurement of 2.5 feet. Measure it again and you get 2.5 feet. Measure it repeatedly and you consistently get a measurement of 2.5 feet. The tape measure yields reliable results.

Validity refers to the extent we are measuring what we hope to measure (and what we think we are measuring). To continue with the example of measuring the piece of wood, a tape measure that has been created with accurate spacing for inches, feet, etc. should yield valid results as well. Measuring this piece of wood with a "good" tape measure should produce a correct measurement of the wood's length.

To apply these concepts to research, we want to use measurement tools that are both reliable and valid. We want questions that yield consistent responses when asked multiple times - this is reliability. Similarly, we want questions that get accurate responses from respondents - this is validity.

Reliability refers to a condition where a measurement process yields consistent scores (given an unchanged measured phenomenon) over repeat measurements. Perhaps the most straightforward way to assess reliability is to ensure that they meet the following three criteria of reliability. Measures that are high in reliability should exhibit all three:

Test-retest Reliability:
When a researcher administers the same measurement tool multiple times - asks the same question, follows the same research procedures, etc. - does he/she obtain consistent results, assuming that there has been no change in whatever he/she is measuring? This is really the simplest method for assessing reliability - when a researcher asks the same person the same question twice ("What's your name?"), does he/she get back the same results both times. If so, the measure has test-retest reliability. Measurement of the piece of wood talked about earlier has high test-retest reliability.

Inter-item reliability:
This is a dimension that applies to cases where multiple items are used to measure a single concept. In such cases, answers to a set of questions designed to measure some single concept (e.g., resuscitation) should be associated with each other.

Interobserver reliability:
Interobserver reliability concerns the extent to which different interviewers or observers using the same measure get equivalent results. If different observers or interviewers use the same instrument to score the same thing, their scores should match.

Validity
To reiterate, validity refers to the extent we are measuring what we hope to

measure (and what we think we are measuring). How to assess the validity of a set of measurements? A valid measure should satisfy four criteria:

Face Validity:
This criterion is an assessment of whether a measure appears, on the face of it, to measure the concept it is intended to measure. This is a very minimum assessment - if a measure cannot satisfy this criterion, then the other criteria are inconsequential.

Content Validity:
Content validity concerns the extent to which a measure adequately represents all facets of a concept. Consider a series of questions that serve as indicators of resuscitation (Heart rate, respiratory rate etc.). If there were other kinds of common behaviors that mark a person as not resuscitated that were not included in the index like cyanosis, pupil size, then the index would have low content validity since it did not adequately represent all facets of the concept.

Criterion-related Validity:
Criterion-related validity applies to instruments than have been developed for usefulness as indicator of specific trait or behavior, either now or in the future. For example, think about the driving test as a social measurement that has pretty good predictive validity. That is to say, an individual's performance on a driving test correlates well with his/her driving ability.

Construct Validity:
But for a many things we want to measure, there is not necessarily a pertinent criterion available. In this case, turn to construct validity, which concerns the extent to which a measure is related to other measures as specified by theory or previous research. Does a measure stack up with other variables the way we expect it to? A good example of this form of validity comes from early self-esteem studies - self-esteem refers to a person's sense of self-worth or self-respect. Clinical observations in psychology had shown that people who had low self-esteem often had depression. Therefore, to establish the construct validity of the self-esteem measure, the researchers showed that those with higher scores on the self-esteem measure had lower depression scores, while those with low self-esteem had higher rates of depression.

Validity and Reliability Compared:

So what is the relationship between validity and reliability? The two do not necessarily go hand-in-hand.

At best, we have a measure that has both high validity and high reliability. It yields consistent results in repeated application and it accurately reflects what we hope to represent.

It is possible to have a measure that has high reliability but low validity - one that is consistent in getting bad information or consistent in missing the mark. It is also possible to have one that has low reliability and low validity - inconsistent and not on target.

Finally, it is not possible to have a measure that has low reliability and high validity - you can't really get at what you want or what you're interested in if your measure fluctuates wildly.

SESSION TWO:

✉Compiled by Anshu:

MEANINGFUL INTERPRETATION OF ASSESSMENT DATA

Validity is never assumed. It is always approached as a hypothesis. It uses theory, logic and scientific methods to collect and assemble data to support or fail to support proposed score interpretations at a given point in time.

For some interpretations of assessment results, only one or two types of evidence may be critical. But an *ideal validation* would include evidence from all four categories: content-related, criterion-related, construct-related and consequences of using.

We are most likely to draw valid inferences from assessment results when we have a full understanding of:

1. The nature of the assessment procedure and specifications that were used in developing it.
2. The relation of the assessment results to significant criterion measures
3. The nature of the psychological characteristics or constructs being assessed and
4. The consequences of using the assessment

Though not always practical, one must gather as much relevant evidence as is feasible with the constraints of the situation. We should also look for various types of evidence when evaluating standardized tests:

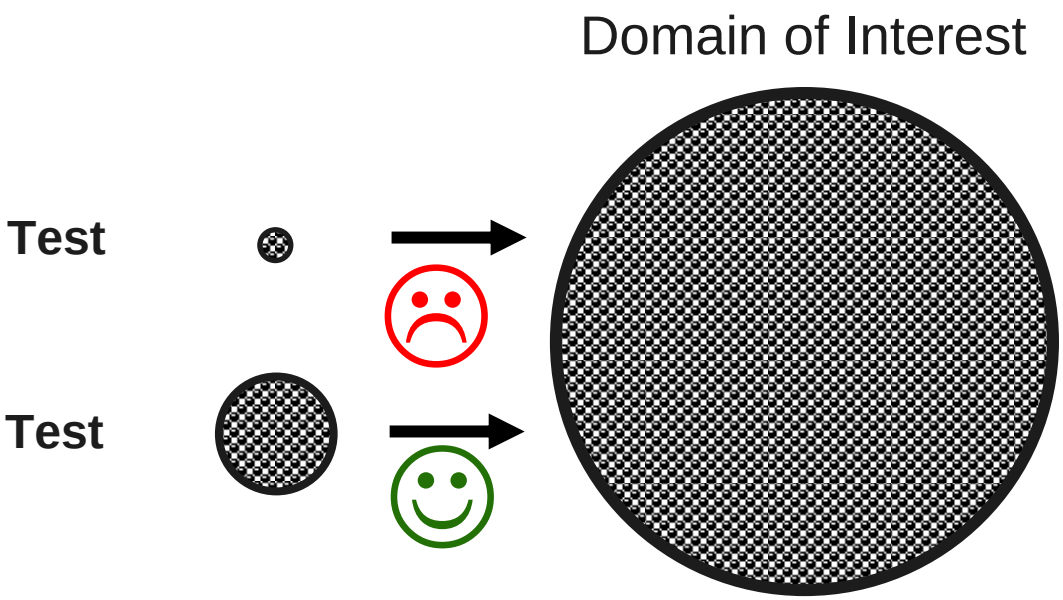
BASIC APPROACHES TO VALIDATION	
Type of evidence	**Questions to be answered**
Content related	How adequately does the sample of assessment tasks represent the domain of tasks to be measured?
Criterion related	How accurately does performance on the assessment (e.g. test) predict future performance (predictive study) or estimate present performance (concurrent study) on some other valued measure called a criterion?
Construct related	How well can performance on the assessment be explained in terms of psychological characteristics?
Consequences	How well did use of the assessment serve the intended purpose (e.g. improve performance) and avoid adverse effects (e.g. poor study habits)?

CONTENT RELATED EVIDENCE:

This is critical when we want to use performance on a set of tasks as evidence of performance on a larger domain of tasks.

E.g. If you wanted to take a test on a certain topic and you used only easy questions or only difficult questions from the topic, the scores would have low validity. The test must be balanced and must be representative of what is expected of the students.

Therefore the key element in content related evidence of validity is the *adequacy of sampling.* An assessment is always a sample of the many tasks that could be included. Content validation is a matter of determining whether the sample of tasks is *representative of the larger domain* of tasks it is supposed to represent.



To be assured that an assessment provides valid results:

1. Identify the learning outcomes to be assessed
2. Prepare a plan that specifies the sample of tasks to be used
3. prepare an assessment procedure that closely fits your list of specifications

Here we are assuming that the assessment was properly prepared, administered and scored. Validity must be 'built-in' during planning and preparation stages by proper administration and scoring

FACTORS THAT LOWER VALIDITY OF ASSESSMENT RESULTS

- Tasks that provide an inadequate sample of achievement to be assessment
- Tasks that do not function as intended due to
  - use of improper types of tasks
  - lack of relevance
  - ambiguity
  - clues
  - bias
  - inappropriate difficulty etc.
- Improper arrangement of tasks and unclear directions
- Too few tasks for the types of interpretation to be made (e.g. interpretation by objective based on few test items)
- Improper administration: inadequate time allowed, poorly controlled conditions
- Judgemental scoring that uses inadequate scoring guides, or objective scoring that contains computational errors.

CRITERION RELATED EVIDENCE:

Two types of studies are used to obtain evidence for criterion related validity:

1. **Predictive study**: It is concerned with the use of test performance to predict future performance on some other valued measure called a criterion. E.g.: Scholastic aptitude test scores maybe used to predict course grades
2. **Concurrent study**: This is concerned with the use of test performance to estimate current performance on some other criterion. With this procedure, both test and criterion measures are obtained at approximately the same time.

Concurrent studies are done for the following reasons:

- To check the results of a newly constructed test against some existing test that has a considerable amount of validity evidence supporting it.
- A brief simple testing procedure maybe substituted for an complex time consuming measure if it provided a satisfactory estimate of study performance
- To determine if a testing procedure has potential as a predictive instrument

In both these studies the relation between test scores and criterion to be predicted or estimated is typically expressed by means of a correlation coefficient or an expectancy table.

CONSTRUCT RELATED EVIDENCE

This evidence focuses on assessment results as a basis for inferring the possession of certain psychological characteristics.

E.g.: A person's reading comprehension, reasoning ability or mechanical aptitude- are all hypothetical qualities or *constructs*, that we assume to explain behaviour.

When we say that someone is 'highly intelligent', a whole series of meanings are associated with it, which indicate how an individual will behave under certain conditions. But before we interpret assessment in terms of these broad descriptions, we must first establish that the constructs actually do account for differences in performances.

Construct related validity for a test includes:

1. a description of the theoretical framework that specifies the nature of the construct to be measured
2. a description of the development of the test and any aspects of measurement that may affect the meaning of the test scores (e.g. test format)
3. the pattern of relationship between the test scores and other significant variables (e.g. high correlations with similar tests and low correlations with tests measuring different constructs) and
4. Any other type of evidence that contributes to the meaning of test scores. (e.g. analyzing the mental process used in responding, determining the predictive effectiveness of the test)

Construct related evidence is the broadest of the three categories and both content related and criterion related evidences are relevant to this category. These help to clarify the meaning of assessment tools.

CONSEQUENCE OF USING ASSESSMENT RESULTS:

Did the assessment

- Improve learning?
- Improve performance?
- Improve self assessment skills?
- Contribute to transfer of learning to related areas?
- Encourage good study habits?
- Encourage independent learning?
- Contribute to a positive attitude?
- Contribute to adverse effects?
    - Lack of motivation
    - Memorization/ rote learning
    - Poor study habits

ESTIMATING RELIABILITY OF SCORES

Would we obtain same results if we used a different sample on the same type of task?
Would we get same results if we took the test on a different day?
Would different raters rate performance similarly?

Reliability of test scores is typically reported by means of a reliability coefficient or a standard error of measurement
Reliability of test scores is lower when the test is short, range of scores is limited, testing conditions are inadequate and scoring is subjective

FACTORS THAT LOWER THE RELIABILITY OF ASSESSMENTS

1. Insufficient number of tasks
   Remedy: Accumulate results from several assessments
2. Poorly structured assessment procedures
   Remedy: Define carefully nature of tasks, conditions for obtaining the assessment and the criteria for scoring and judging the results.
3. Dimensions of performance are specific to the tasks
   Remedy: Increase generalizability of performance by selecting tasks that have dimensions like those in similar tasks
4. Inadequate scoring guides for judgemental scoring
   Remedy: Using scoring rubrics or rating scales that specifically describe the criteria and levels of quality
5. Scoring judgements that are influenced by personal bias
   Remedy: Check scores with those of an independent judge. Receive training in scoring and rating if possible

E.g. *(Swanson, 1987)* Reliability of an oral examination

Testing time in hours	No. of cases used	Same examiner for all cases	New examiner for each case	Two new examiners for each case
1	2	0.31	0.50	0.61
2	4	0.47	0.69	0.76
4	8	0.47	0.82	0.86
8	12	0.48	0.90	0.93

As is evident, adequate reliability requires substantial sampling (therefore resources: testing time, examiners, patients, etc.)

DISCUSSION TOPICS FOR WEEK TWO:

We now invite your opinions and views on the following issues:

1. What are the deficiencies which you feel affect our present day assessment methods?
2. What are the factors which lower validity and reliability of :
    - Multiple choice questions
    - Essay type questions
    - Clinical examination
    - Oral examination/ viva voce
3. What are the sources of validity of the assessment tools you plan to use in your curriculum innovation projects?

COMPILED RESPONSES FROM FELLOWS AND FACULTY

The need for validity evidence

✉ **Bill Burdick:**

We cannot always have our ideal test results that give us the exact same result with repeated measures and that mean precisely what we say they mean. We must, however, have a test result with at least modest measurement stability for it to have any meaning at all. For example, consider the statistics test result above. Assume I have accumulated the validity evidence described – so far, it sounds like we have a valid exam. But what if the teacher obtained the test score by randomly selecting a score, and that score happened to correlate with my other measures of ability? It would mean nothing, since the next time I took this test, and my teacher randomly selected by score, it is not likely that it would correlate again.

In other words, a reliable test result is a prerequisite to a valid test result. One can have a reliable test score that is not valid (example of shoe size); but one cannot have a valid test result that is not reliable.

Sometimes a test score does not need to be highly reliable to be useful. For example, test results used for formative feedback might not be highly reproducible, but they could still be valuable to the learner. (Of course it they were as unreliable as pulling numbers randomly, then the feedback would have no meaning either).

However, if there is no evidence to support the conclusions we make from a test result, we must question the validity of that result. A multiple choice test used to make an inference about qualifications for clinical post might be an example – highly reliable, but evidence to support the inference might be hard to find.

The focus is often on reliability since that is an easier concept to boil down to a number. Validity of a test result is often demonstrated after a collection of research studies, and even then, there will probably be room for debate.

Lots of terms are used to describe the different types of evidence for claiming the validity of a test result for a particular inference. The terms have been used in different ways over the years by different authors. More important that the terms, is knowing how to look for validity evidence. Does the score correlate with other measures of the same domain? Does the score predict future performance? Does the score correlate with other domains within the same test? Does it negatively correlate with scores that indicate opposite skills? Do the score results make sense when one simply looks at them? What impact on student behavior has the test had? Each of these questions relates to different kinds of validity evidence (specifically: concurrent validity, predictive validity, construct validity, discordant validity, face validity, consequential validity). The labels mean less than the questions.

Which Type of Validity is More Important?

✉ **Tejinder Singh**: More than content validity, I personally feel that predictive validity (does PG entrance result mean that a student is better postgraduate than other) and consequential validity (what effect a test has on the learning habits) are very important in shaping learning behavior.

When we introduce a new examination method, everyone gets impressed, we get very nice and encouraging feedback from students and faculty, it even gets published- however, the real proof lies in predicting the future performance of the graduate vis-a-vis scores in the tests. With our present understanding, it is difficult to quantify the validity aspect in numerical terms. (similar to output v/s outcome concept).

It may be of interest to you if I share my experiences with our old students. We get a large number of requests from various boards and bodies in USA for verification of our graduates in that country. When we take out the old records, it often surprises me that students who are doing very well now, had no impressive records as students. In fact, many of them took more than one chance to pass their examinations. To me, it only suggests that our system fails to predict the future performance very dismally.

✉ **Sanjay Bedi:** Even within our own country during class reunions I do not see any correlation between professional success and marks/ positions during undergraduate period. You cannot generalize. Many other factors/skills come into play.

✉ **Sheena Singh:** That students are being prepared very well, is an interesting thought. I agree that the grades obtained during undergraduate studies are not always predictive of future success.

But as we get better acquainted with the milieu and mature in our chosen field, and if we get an optimally stimulating environment plus other motivating factors of psychosocial nature...many of us become successful later on.

✉ **Anshu:** As far as predictive validity is concerned I agree our tests do not measure the traits that count- leadership, commitment, attitude and aptitude.

✉**Stewart Mennin**: It could well be that the current assessments have weak evidence for predictive validity. It could also be that students are being prepared very well to be stimulated further and grow and develop further along in their practices I always had medium to low grades in school even though I studied very hard. It wasn't until I was in my doctoral and post- doctoral programs that I began to develop better understanding. My early grades were not predictive.

✉ **Sanjay Bedi:** What about gut feeling? Is it valid enough a little bit reliable too?

✉ **Vivek Saoji:** Gut feeling or sixth sense, scientifically termed as "face validity" is no validity and quoting from another of Steven Downing's articles: "The term face validity is sloppy at best and fraudulent and misleading at worst". So there is no place for face validity in any meaningful interpretation of assessment data.

Also quoting from the same article: "Face validity is a "garbage can term", a term without much real meaning or with an inconsistent and confused meaning that adds little or nothing to our understanding of assessment data".

How valid is valid enough?

✉ **Anshu** Whether in research or in assessment, choosing a feasible tool is one of the most difficult decisions to make.

In research for example- if one wanted to investigate the dietary intake of a substance- what would the ideal method be? One would have to weigh and make duplicate helpings of everything a person ate daily for a few days and have it analyzed biochemically. That would be a very cumbersome and impossible task to undertake. So we resort to a 24 hour recall where we ask the person what he ate in the last 24 hours. It is not an accurate method, but is far more acceptable and feasible.

Similarly, the ideal method in assessment would be to test all the knowledge, skills and aptitude that is expected of a medical graduate... a very far fetched ambition. What we therefore rely on - is proper sampling of what we need to test, and then choose the most representative test of all.

The issue at the practical level, is therefore, not which is the valid method, but what is the size of the error in adopting that feasible option. And by implication, does an error of this

size alter the interpretation of our results? When impaired validity leads to misclassification, how much of a difference does it make in the intended situation.

Role of Statistical analysis in validity and reliability

✉**Tejinder Singh:** In our book, there is a chapter on various statistical methods used for calculating reliability. Validity is difficult to be mathematically proved.

✉**Payal Bansal:** The Principles of Medical Education has basic formulae for calculation of reliability. Other analyses used for interpretation of assessment data include a variety of descriptive statistics, correlational analysis, student t test, ANOVA, etc. Basically statistical methods used in social sciences and psychology are used in educational research. Talk to a statistics professional at your institution for good resources. "Statistics - a spectator sport" - by Jaeger (Sage publications) is my favourite - it has no formulae only explanation of what each test means, with examples.

✉ **Anshu:** When we first started this online discussion, the faculty felt we should veer clear of actual statistics, because most of us were statistics phobic. I am also one of those people who doesn't like being tangled in a web of numbers.

But for those who are interested, do take a look at this site.
http://onlinestatbook.com/rvls/

It is interesting not only for the simple text (the Hyperstat link) but for the simulations of the statistical methods which most of us struggle to understand.

I am also attaching an article from the BMJ on Cronbach's alpha if you are interested.

Sensitivity/ Specificity Vs Validity/ Reliability

✉ **Monika Sharma, Aroma Oberoi**: Can we compare the concept of reliability and validity with the concepts of sensitivity and specificity?

✉ **Tejinder Singh:** I would rather talk of sensitivity/specificity in the context of a lab test.

✉ **Anshu:** When validity tests are applied in assessing the performance of diagnostic and screening tests, we use two indicators- sensitivity and specificity.

In this situation, the objective is to evaluate the ability of a test to distinguish between true disease-positive and true negative individuals. Sensitivity is the proportion of subjects who truly have the disease (true disease-positive). Specificity is the proportion of individuals who truly do not have the disease (true disease-negative).

Most often, in a diagnostic test- the results are dichotomous- either positive or negative. Which makes calculating sensitivity and sensitivity fairly simple. However, in cases where the results of a test are **on a continuous scale**, **we need to decide a cut-off.**

That cut-off limit must be defined in such a manner, that the sensitivity and specificity is optimised **for the purpose intended**.

E.g. A screening test is done for a substance in urine to detect a disease. A study reveals that the substance is present from 5 mg to 98 mg in the population. There is overlap between distribution of concentration of that substance with and without that disease. If I use a cut off at the extremes- like 5 mg- all of the population will be labelled as disease positive. If I use a cut-off limit of 98, clearly all people will be labelled disease negative. Clearly, both cut-offs do not serve their purpose. The challenge is to find a cut-off value between the extremes which satisfactorily optimises the values of sensitivity and specificity.

In terms of assessment in medical education, we scarcely have such strict dichotomous criteria as in diagnostic tests. The results are dichotomous (pass-fail), but the basis of deciding cut-offs is vague. When we define a pass percentage as 35% or 50%, do our criterion consider the difficulty of the paper, the content or the reason for which the test is being given?

Let's take the example of an exam for passing from a medical school. Here we do have a dichotomous result as pass or fail. But does our test cut-off decide whether all students who pass MBBS are actually capable of working independently? Do our assessment tools guarantee that a student who has been certified pass by us will not kill someone later by his lack of knowledge or negligence? The concept of 'specific' comes in here- does our exam specifically filter out those students who are incapable of practicing independently?

Let's consider Dr TS's favourite example of a selection exam for choosing students for a medical school. Strangely, here we do not use marks as a criterion at all. The criterion here is the number of seats we have in our school. So, we choose the top students who bother to give that exam. It does not mean that the others who were not selected were not capable enough to do the same course. The key issue for us as examiners is to see that the exam is 'specific' enough to filter out those students who are not capable enough to pursue that course. Even if one 'wrong' student is selected, your exam methodology is dubious. But do our exams even test things other than knowledge? Aren't aptitude, the right psyche and motivation to pursue medicine equally important?

So then how do we determine our cut-offs in assessment?

✉ **Payal Bansal:** You have raised a very pertinent issue about cut-offs. Standard setting is an important component of any assessment model, and there is a lot of literature if you would like to explore.

✉ **Sheena Singh:** To add to the argument, students are very varied individuals and my study on motivation made me realize that diversity is the norm.

We want to standardize the medical education process as much as possible and have uniform criteria for qualifying all students to ensure clarity of instruction and fairness in evaluation ...hence reliable and valid testing tools (sensitive & specific)

Medicine is both a science and an art and there is only so much we can do. However there is still a grey area of subjectivity which we can perceive about the student taking an examination during viva voce or interview for post graduate entrance that gives an experienced teacher a 'gut feeling' about the ability of the student.

It may seem contradictory to all principles of medical education to state this, but I don't mind risking it to have an opinion on this.

We are in an era of evidence based medicine, hence we should seek evidence from long term follow up on the competence of the doctor in real life practice and how closely it correlates with his/her performance in objective assessments in qualifying exams.

SESSION THREE:

ASSESSMENT TOOLS: STRENGTHS AND WEAKNESSES

Deficiencies in our Present Evaluation System

✉**Monika Sharma:** Reliability is the ability of a test to be reproducible with the same or nearly the same results every time. In other words reliability is consistency. MCQs are fairly reliable in testing knowledge aspect and valid too. Take another example, questionnaires used for evaluating a programme. How reliable are they? Various external influences such as the observer's state of mind would affect his responses. I feel they are still valid.

Validity is the ability of a test to be able to give a reliable opinion about what it is intended to test or validate. Direct observations of performances and attitudes of students, working in a team or in community postings is would be a valid indicator of their ability to work in a team, work in the community, because it is testing the skill it is intended to test But is it reliable? Students' behaviours may keep changing and hence the evaluation results would keep changing with different observers.

Let's take some more examples. Taking anthropometry of a child to evaluate the nutritional status is valid, but not reliable as there could be some inter-observer variation. Thesis writing during postgraduation, is intended to help the student learn how to do research, write an article and get to know statistics. Is it reliable? Not at all, ten students would produce the same thesis very differently. Is it valid? I would say no again as how many students do the thesis themselves in toto?

At the end of it all, I would consider both reliability and validity to be important properties of an evaluation tool, an unreliable valid test would be as useless as an invalid reliable test.

✉ **Anshu:** Whether an assessment tool is valid or not depends on the purpose for which it is used. If a dissertation is used as an assessment tool to gauge if a student has understood basic principles of research methods, I do think it is a valid tool. Inspite of a student's working under supervision. But if your basic premise is that after completing thesis, the student must be able to independently conduct research studies, more clauses need to be added to how he/she does the thesis.

You say long answer questions are not valid for evaluating knowledge. Are they a useful assessment method to evaluate the ability of a student to reason and critically organize his thoughts? Is there no place for these questions in medical education? Can we improve the format and thus the reliability of this format? Is the problem with the tool itself or with the scoring pattern?

✉ **Varughese Varkey:** You brought out the point whether all students who pass out MBBS are capable of working independently? They are not. It is probably the failure of our training process. How much independence do we give to our interns in handling patients? They record histories, collect reports etc. Post graduate residents are the ones who manage the patients under the guidance of consultants. Interns are just bystanders. How can we expect them to handle a patient independently on a fine morning?

I do agree with you that we test only their knowledge for the exams. We don't have anything to test their professionalism.

✉ **Anshu:** I feel we need to move away from marks and focus on *competency based assessment*. If a student can do ten procedures right, which is expected of him- he should clear the exam. We need to focus on log books. And they ought to be evaluated seriously, not like the joke they are usually turned into here.

I was never taught to resuscitate a patient as an undergraduate. I've never been taught to intubate a patient. Something I'm still ashamed off. But that's where we need to shake up our system.

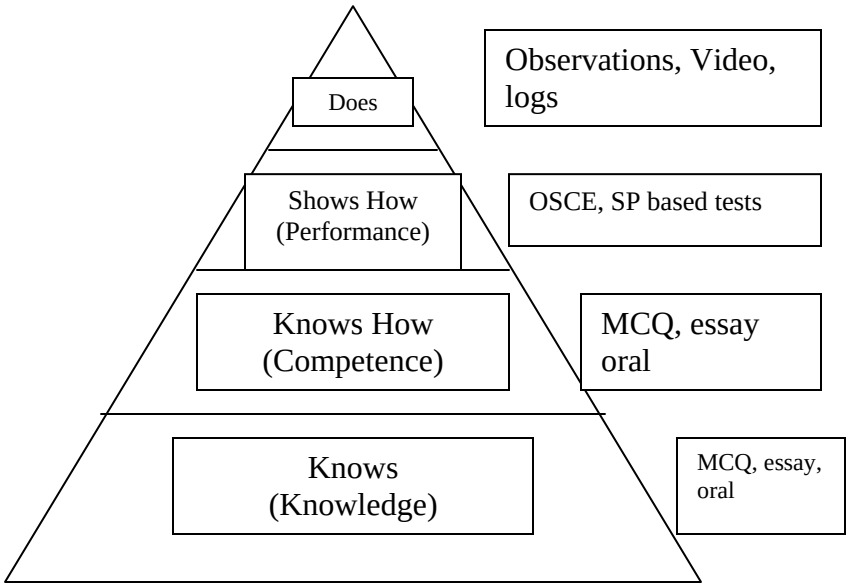
✉ **Monika Sharma:** I do consider viva/oral exams as the only existing evaluation tools being used now-a-days for assessing the independent capabilities of the student and am totally for it. We have other tools that test knowledge: OSCEs, theory papers etc. while the same can be used for both. E.g. properly framed OSCEs can test the student's judging capabilities, affective domains as well as psychomotor, structured short questions or even structured long questions can be used to do the same. It is unfortunate that all available tools are being used for checking knowledge-not that they cannot..
It is the lack of knowledge of the people preparing questions to use them properly for their meant purposes. This brings us to the fact that we need more teacher training sessions and opportunities to bring about a change.

✉**Rita Sood**: Just came across this very simple and useful article on role of assessments in higher education

Miller's Pyramid

✉ **Vivek Saoji:** It is always a challenge to identify the right methods to assess the different levels of competence, which are not only able to assess the examinees in all the domains but have good psychometric properties so to say that they are valid and reliable.
Miller in 1990 developed a simple model "Millers Pyramid" which defines the level of competence and the appropriate assessment method for each of these levels.
This model will be useful when you are planning the assessment and help you identify the appropriate method.

Millers Pyramid (Miller 1990)



✉**Sanjay Bedi:** Selecting assessment tools to match your needs

✉**Stewart Mennin**: I hope that in the discussion about assessment does not lose sight of a continuous consideration in our minds about how we are choreographing the learning/teaching activities so that they are consistent with the assessments? Are students experiencing learning activities that provide opportunities for them to demonstrate successfully their learning on the selected assessments? Some would say that this is obvious, others may not.
✉**Chandrika Rao:**

Two principles to remember:

- The success of any skill based assessment depends on finding a suitable balance between validity and reliability and between the ideal and the practical.
- When pass-fail decisions are being made, a skill based assessment should be "criterion referenced" (that is, trainees should be assessed relative to performance standards rather than to each other or to a reference group)

Viva Voce

✉ **Anshu:** I feel the oral examination is the only place we assess a student's communication skills, his ability to defend his decisions and his presence of mind- all essential attributes of a medical graduate who is about to be let loose on society.

And there are ways of making viva more structured and reliable. Our present exam system is full of loopholes because our examiners are untrained, not reliable. A standardized examination can work wonders.

Handout on viva voce

✉**Chetna Desai:** **The debate over viva voce**

The term viva is used generically for an oral examination. It may be used as apart of the formal general assessment or to assess a project or dissertation (Presentation/ demonstration of the project work). Apart from the conventional **advantages** of the viva discussed earlier here are a few more to reflect on:

- Viva helps to identify and assess the marker candidates among a cohort and assess the overall standard of the course.
- A properly conducted viva can effectively assess plagiarism in cases of projects or research work.
- It may help borderline students cross the line.
- Allows feedback from students for improvement and gives an opportunity to students to express their views on the fairness of assessment by the examiners.

Drawbacks:

- The major drawback is the lack of fairness, transparency and reliability in the process of conducting the viva.
- It often tends to get too subjective
- Uniform standard of assessment is not maintained; some students may hence be judged in an unfair manner.
- Often neither the teacher nor the students are trained in handling the viva.
- Language barrier and lack of articulation is another hurdle that the student may face.

A few suggestions to overcome these drawbacks:

- Viva should only be used to award grades/ for classification and progression of students. Individual marks to the students should be avoided.
- At least 2 examiners should undertake the viva. This will reduce the error.
- The format, duration and the subject to be discussed should be notified to the examinee and the examiner beforehand. The proceedings should be open, consistently applied, documented and records should be accessible.
- The questions asked should be specific to the subject and not wide-ranging.
- Avoid closed, hypothetical, leading questions.
- Examiners should talk as little as possible.
- Avoid using value judgment words such as Good etc, as they are likely to be misinterpreted by the candidate.
- Do not use patronizing language or inappropriate body language.
- Be sensitive to the language barriers and possible nervousness and distress of the student.
- Do not make race, gender or background assumptions when making comments or asking questions.

- Academic judgments in viva can be influenced by extraneous factors such as appearance; first impression, contracts between two candidates, pleasing or abrasive personality of the candidate etc. hence beware of such influences.

A viva methodically conducted is an important assessment tool and has many advantages that the conventional written examination does not have. The viva can be assessed on rating scale, through role-play and by critique of the video of an oral examination.

Viva voce Vs OSCE

✉**Tejinder Singh:** One of the teachers at CMC had devised this system for viva, which is worth trying-
Collect questions keeping in mind the learning objectives etc. and divide them into 3 categories as simple, moderate and difficult. Have each question printed on a different card, with easy on green, moderate on yellow and difficult on red. Let the student pick up 2 cards from green and answer the. If he does that well, let him pick 2 from yellow and then let him progress to pink. This allows unbiased questions and helps to move from easy to difficult.
The key to success of the system is having a large number of questions.
✉ **Chetna Desai:** Seems to be a nice method. Tedious maybe, but could be effective.
How do we decide the difficulty level of the questions? Based on teacher's perceptions or on students performance on these questions in the previous exams?

✉**Mrunal Ketkar:** How is the assessment done after picking up and answering the questions? Is there a different cut off depending on the difficulty level? If so then it really means a lot of efforts to be taken by the faculty.

I think unconsciously all of us are doing this kind of grading of questions in our mind when we are taking a viva of students. One thing that always nagged me is the personal bias that may unconsciously occur. Let me explain we observe the students right from their 2nd year. Sometimes a student, whom we know has been good through out, may not be doing well in a particular viva. Will his overall performance for the year be affected by one viva?

✉**Sanjay Bedi:** A Learning Management System will be helpful in this scenario. The difficulty level of each question can be found out statistically depending upon the percentage of students who answer it correctly and then grading allotted accordingly.

Also see the links below: http://en.wikipedia.org/wiki/Learning_Management_System

http://www.google.com/search?sourceid=gmail&q=Learning%20Management%20System

✉ **Anshu:** Someone I know also uses a similar method where questions of the same level of difficulty are piled into separate boxes and the student opens a chit, reads the question and answers it. The problem is with us- the examiners- who hate it when they have to play a passive role! Being a good listener is a very tough job.

I personally think the OSCE is one of the most boring exams to take as an examiner - too mechanical. I enjoy taking vivas- the sad part is that most of the enjoyment is at the expense of the poor candidate.
✉ **Tejinder Singh:** I think you are right. Assessment deserves as much attention, patience and thought as any other research. Most of the times, we, as examiners, are not willing to do that.

✉**Monika Sharma**: In contrast Anshu, if given a choice between theory and OSCEs, I would prefer to do the second one. Though the inputs are higher initially, OSCEs are easier to assess. Theory papers of the present day type are too subjective and checking them can be really boring and sometimes irritating considering the mixed bag of students in every class.

✉**Venugopal Rao:** The system followed in CMC has lot more advantages as you mentioned. In fact we can try it in PG exams also. This is non biased and students get equal chance of scoring marks. But as you rightly pointed it needs more number of questions and lot more patience to prepare them.

✉**Vivek Saoji:** Viva Voce or the oral examination is an integral part of our examination system and so far you have very rightly pointed out the advantages and disadvantages of this method.
To get further understanding and pros and cons of oral examination I am attaching a very interesting and very useful paper on oral exams. I thought this is a very short paper explaining the concept in a very simple language and will definitely contribute in the existing discussion in getting the concepts right.

✉**Mrunal Ketkar:** I agree that taking viva is much fun. But sticking to the principles of our discussion don't you think it is lower down on the scale of Reliability and Validity.

Setting up a OSCE is lot of hard work initially but quite relevant especially during the initial clinical term ending exams, when we are teaching them the must know things like History taking, examination of swelling or an ulcer. And once you have a check list made, it is more or less permanent and your next OSCE assessment is faster and easy.

✉ **Payal Bansal:** I think OSCE allows for wider sampling, and, like MCQ's, if well constructed can measure complex learning. It is well accepted and widely used the world over for clinical skill assessment. Of course, the assessment can be a composite one, with OSCE being one of the components

Should oral practice examinations be used for formative or as high-stake evaluations?

✉**Chandrika Rao:** Should oral practice examinations be used for formative or as high-stake evaluations?

✉**Anshu:** I don't see why a well organized objective structured oral examination cannot be used for a high stakes exam. The only problem- and it is a gigantic problem- is with the logistics. Just imagine having to conduct the Pre-medical entrance examination of nearly lakhs of students each year using this format. It is simply not feasible- given the time, resources, trained examiners and volunteers etc. In formative assessments, again it depends on the student-teacher ratio plus time. With the constraints of time in our present curriculum- it is not too easy. If you have the money and the resources- go ahead.

Where these exams can be used- are in the second stage- after you have screened one group for knowledge... and now want to test their psychological attributes and aptitude etc. for the profession. Something like your UPSC exam for IAS where you have a two stage exam.

It is important to test for communication skills and personality traits depending on what you want. Would you want to appoint a professor in your department on the basis of a CV without a one to one interview?

I recently gave the IELTS exam, and I returned *maha*-impressed with the professionalism of the British Council. But they had more than a dozen examiners for 50 odd students. The speaking test was conducted at different times in three rooms throughout the day. Each interview was recorded faithfully. They paid attention to politeness, hunger pangs, comfort and a dozen other things we never bother to think about- actual VIP treatment.

But as the examiners were compelled to read out questions from their standard papers, I could see fatigue on their faces. They were being gauged and assessed too.

How does one test the ability of a student to carry out self directed learning?

✉**Chetna Desai**: What are appropriate tools to assess self directed learning? How would one test "ability to carry out health education activities in the community" or "ability to work in a team"?

✉**Payal Bansal:** The first thing that came to my mind when I read Anshu's question on self-directed learning was the Portfolio. Portfolios are not used in India in medical education as yet, but are widely used elsewhere, for learning as well as assessment of

learning. An essential component of portfolios are reflections by the learner on specific tasks/problems allotted as well as their own learning.

Concept maps and Gantt charts which you are using right now in your projects can be other tools - independently or maybe as part of a portfolio. I did a quicksearch on the net and found this link which I liked.
http://home.twcny.rr.com/hiemstra/sdltools.html

For "ability to carry out health education activities in a community", projects would be a good tool, if criteria for assessment are well designed. Other methods could be direct observation in real life situations (village posting, interaction with families) or simulated situations.

For assessing team ability - direct observation again, based on a group dynamics guidelines could be one way. Self and peer assessment are also used.

Have you re-visited the assessment toolbox? It might be worthwhile to do so at this point.

✉**Mrunal Ketkar:** Group Discussions observed by the faculty (as usually done as a part of the selection criteria in management world) could be one of the answers.

✉ **Monika Sharma**: I feel it would be difficult to comment on the ability of a student in relation to the community and as part of a team by a single or one point assessment tool. A continuous observation of attitudes and behaviour, as part of internal evaluation by concerned teachers/supervisors or several formative evaluations or even feedback from sources other than the students themselves would be more useful.

Clinical Examination / Long Case:

✉**Chandrika Rao:**

1. The long case has been an essential tool in the assessment of clinical skills
2. It evaluates performance with real patients and enables students to gather information and develop treatment plans under realistic conditions
3. Problems have been found with the reproducibility of scores generated by the long case, less reliable. Implicit in the use of the long case is the assumption that if the student was examined again with another patient and different examiners, the results would be the same.

Three major factors explain why the long case has problems with reproducibility. In decreasing order of importance they are

1. Case specificity of problem solving,
2. Differences between examiners,
3. Variability in the aspects of an encounter evaluated.

Modifications to the long case to improve validity and reliability are:

1. Increasing the number of student-patient encounters
2. Increasing the number of examiners
3. Increasing the number of aspects of a competence assessed and standardizing them across examiners has a modest positive effect on the reproducibility of scores.

An interesting article: The death of the long case?

✉**Avinash Supe:** Regarding reliability and validity of clinical examination - have a look at this slide:

ASSESSMENT OF CLINICAL SKILLS

Exam	valid	reliable	generalisable
Long case	++	-	-
Short case	+	-	+/-
Observed long case	++	+	-
OSLER	++	++	-
OSCE	+/-	++	++

✉**Avinash Supe:** Reliability and validity – Skill assessment

The reliability of a test describes the degree to which the test consistently measures what it is supposed to measure. The more reliable a test, the more likely it is that a similar result will be obtained if the test is re-administered. Reliability is sensitive to the length of the test, the station or item discrimination, and the heterogeneity of the cohort of candidates. Standardized patients' portrayals, patients' behaviour, examiners' behaviour, and administrative variables also affect reliability.

Factors leading to lower reliability

- Too few stations or too little testing time
- Checklists or items that don't discriminate (that is, are too easy or too hard)
- Unreliable patients or inconsistent portrayals by standardized patients
- Examiners who score idiosyncratically
- Administrative problems

The validity of a test is a measure of the degree to which the test actually measures what it is supposed to measure. Validity is a property of test scores and justifies their interpretation for a specific purpose. The most basic evidence of validity comes from documenting the links between the content of the assessment and the curriculum's objectives and from the qualifications of those who develop the assessment.

Questions to ensure validity

- Are the patient problems relevant and important to the curriculum?
- Will the stations assess skills that have been taught?
- Have content experts (generalists and specialists) reviewed the stations?

Multiple Choice Questions (MCQs)

✉**Anshu:** Are MCQs valid? All they manage to test is the cognitive domain. In our present system, how many of us bother to sit down and make problem solving or higher understanding/ interpretation testing items. Any one who has bothered to sit down and make a question paper would know that the easiest questions to make involve pure recall.

How much importance needs to be given to training of paper setters? The agreement on what constitutes a correct answer is often a bone of contention amongst colleagues.

✉ **Mrunal Ketkar:** The one difference that came to my mind between undergraduates and postgraduates is the importance given to the ability of a postgraduate to think independently about a clinical problem, apply his/her knowledge by reading and that of experience, to demonstrate a correct logical way of thinking and then arrive at a decision .The decision need not be always be right on i.e. many times we encourage them to give a Differential Diagnosis and ask them to defend it. I think to construct a MCQ which checks all these abilities is little difficult.

✉**Chetna Desai:** Here is a link that will help**:** nbme_guide.pdf Also please refer to the attachment guidelines for writing MCQs

✉**Monika Sharma**: Dr Tejinder's book also has a simple and elaborate chapter on MCQs and evaluation of MCQs.

MCQs as assessment tools are widely liked for being specific in the item being tested- usually the cognitive domain, which is the only drawback, I feel. Most of our previous year's projects have proved that they are difficult to prepare and need some training. They need to be validated before being used, which is what most of us don't do.
Dr Varughese's project on MCQs last year in the pediatric department showed that though there were almost 9 faculty members with some formal introduction to preparing MCQs, only 20% of those prepared were considered useful. All the others had to be discarded.

I would really like to commend the people who write books on MCQs for entrance tests. The effort must be too tiring. Are all of them really validated? If not aren't the entrance tests facing a question mark on their reliability and validity??

Considering the hard work involved, how many of us would like to use them frequently. The assessment may be effortless, but the inputs are deterring.

✉**Sanjay Bedi**: What makes a good MCQ?
http://www.tlc.murdoch.edu.au/eddev/evaluation/mcq/mctests.html#strengths
See this link too
✉**Venugopal Rao:** Does MCQ have a role in the summative evaluation of postgraduate learning

✉**Chetna Desai:** What are the advantages and disadvantages of MCQs especially as tools for summative assessment For PGs? To begin with: MCQs are good tools for formative evaluation if carried out on a regular basis. They help identify the learning areas that are

- Deficient
- Difficult to understand
- Misunderstood by students
- Need more emphasis or alternative approaches in teaching
- Also identify students who need personal attention (poor performers)

How good they are as tools for summative evaluation? They have certain drawbacks:

- Not enough as stand alone assessment tools. Evaluation of PG students needs a multipronged approach.
- Single response MCQs often assess only recall. Higher levels of MCQs are necessary to evaluate analytical and problem solving skills. These are difficult to frame and time consuming.
- A large question bank in which each item is analyzed for difficulty and discrimination indices is necessary to avoid repetition.

- The problem of copying and secrecy is marked with MCQs as compared to other assessment tools.
- It has been often observed that students who do well in Essay and SAQ based exams fare not so well in MCQ based tests. This is often evident in the various entrance examinations that follow the qualifying exams, be it for medicine, engineering etc.

✉**Tejinder Singh**: You are right that students who otherwise score well do not do well on MCQs. It is a very interesting and consistent phenomenon- more so as my son, who stood second in the district in 10+2, was at 200 rank in the PMET. Though it is not backed by any study or other authentic evidence, I have the following hypotheses-

- Most of the times, the number of MCQs is 'limited' and if a student has read many books on MCQs, the chances are that he will have encountered that question earlier. Students learn to recognize the correct answer, rather than knowing the correct answer. Students scoring higher in entrance examinations often drop a year or do not go to classes after September to prepare for the entrance tests.
- The formulation of MCQs is faulty, allowing for intelligent guesswork.
- In general, a faulty MCQ has more problems for a bright student as he tends to read too much into the stem.
- Answering MCQs is a skill in itself, totally unrelated to knowledge. On the net, there are a number of sites on how to hack the MCQ! I once took an Engineering MCQ test just to test on this point and scored- any guesses- 76% marks, without knowing anything about the subject.
- The classroom model does not make use of MCQ as a evaluation tool in most places and for many students, it is in fact the first encounter with a very high stake MCQ test.

The quality of MCQs in available books always remains a problem. We published a study on this issue-

Y.K. Sarin, M.V. Natu, A.G. Thomas, Tejinder Singh .Item analysis of published MCQ Indian Pediatrics Nov. 1998 Vol. 35 Page 1103 YK Sarin
http://www.indianpediatrics.net/nov1998/nov-1103-1105.htm

Comments to YK Sarin
http://www.indianpediatrics.net/may1999/may-523-524.htm

✉**Chetna Desai:** That's interesting but true! I had a colleague during my residency who had cleared his PMET...He had once said-if you do not know the answer to an MCQ and just want to make a guess --answer "C" and you are more likely to be right. I do not know whether it was his personal observation/hearsay/wild guess or?!
But it certainly brings us back to square one. How good are MCQs as assessment tools in such competitive tests which decide the future careers?

✉ **Varughese Varkey:** MCQ is a good tool in student assessment, if used properly. Most of the MCQs that we use are of the simple recall type which test the only the memory of the student. Higher forms of MCQs can be constructed which are of application of knowledge or evaluation of knowledge type in which the student needs to

analyze and apply his knowledge to answer a certain question. These are of course difficult and time consuming to construct.

Apart from construction of good MCQs there is another aspect to it – to validate these MCQs by test and item analysis. Only after validation these can be used for student assessment. I am not sure how many colleges in India are using validated MCQs for admission tests both for under graduation and post graduation.

✉ **Chandrika Rao:** I came across an article on MCQ which is easy to read. Examples are non-medical, still I felt it is very clear.
Validity and reliability are considered at various levels:
Experience is a hard teacher because she gives the test first, the lesson afterward." -- Vernon Law

"The quality of the assessment can only be considered in the context of the purpose." Jay Parkes, Ph.D.
It basically depends on what we trying to assess.
Assessment can also be categorized according to purpose.
1. This list of assessment purposes was compiled from Foxman (2000), Leat & Nichols (2000) and Heady (2000):

- Motivate the student
- Diagnose student understanding and leaning needs
- Teach the student
- Quantify student achievement for reporting to parents and others (perhaps related to 1)
- evaluate the pedagogy
- evaluate the teacher
- evaluate the school
- evaluate the school district (or other organizational unit)
- pass or fail the student to the next "grade"
- accept a student into or reject student from a school or program

or the one known the most:
2. Kirkpatrick's model of assessment:

- Level I: Evaluate Reaction
- Level II: Evaluate Learning
- Level III: Evaluate Behavior
- Level IV: Evaluate Results
- Fifth level was later "added" for Return On Investment

MCQs can be assessed on these points.
Assessing by multiple choice questions

✉**Payal Bansal:** Literature suggests that multiple choice questions can **reliably** measure upto the **first four cognitive levels of Bloom's taxonomy i.e. knowledge, comprehension, application and analysis.** However,major deficiencies exist in our item writing process, because of which we fail to use MCQ's optimally.

For example, the majority of faculty has not undergone any training. A few may have attended some session or workshop, which give only a superficial understanding of the process. No guidelines are provided to the item-writers when they are assigned the task of developing items. The items are constructed on "chapters" or "topics" and there is no clear-cut exam blueprint. If the first language of faculty is not English, the chances of grammatical errors are significant. The quality of the stem is also affected. Negative items and lower order items, being easy to write, are often constructed. It is therefore not surprising to have many flawed items.

Downing and Haladyna(2002), have identified 31 guidelines which should be taken into account while constructing an MCQ.
It has also been said in literature that it a "good" MCQ takes at least ONE HOUR to write!
Below is a link for a good reference on MCQ writing and item flaws-
http://radiographics.rsnajnls.org/cgi/content/full/26/2/543

MCQs are a good tool, but we have not utilized them to their fullest potential. We need to educate ourselves more on the item writing process. What is heartening to know, is that, we are not alone - people evrywhere have these challenges!!
Use the search words "item writing flaws" to look for more literature.

It is also cited in literature that C is the most often chosen correct response by the item writer. It is important as part of item writing process, to evenly distribute the correct options between A, B, C and D.

Good MCQ's have to be made "guess-proof" by paying attention to all these details.

✉ **Monika Sharma:**
http://tep.uoregon.edu/resources/assessment
This is a link to the site of oregon university on teaching effectiveness program. It not only provides simple description of how to write MCQs, there are several other issues dealt well-such as how to teach in a large classroom, how to evaluate students in a classroom, preparing to teach, presenting facts, motivating students, grading and managing classes...and even balancing one's life.
I think it should be of some help.
Here's an article on examples of classroom assessment.

✉ **Stewart Mennin**: The effectiveness of MCQs for PGs would depend on the context in which they were embedded. For example, if there were rich contextual situations, cases, patient problems, community issues, etc. in which complex issues unfolded and then there were MCQs that required insight, understanding and choice on the part of the PGs; to recognize situations and choose applications of information in the associated

context and setting, it would be an effective application of this assessment tool. Of course, it requires ability to write such questions. Here, perhaps, is a question of faculty development in assessment. What would a program to develop such questions look like? How would you establish validity and reliability?
✉**Avinash Supe:** This is a very important point - MCQs for PGs

- Medical Council in 2000 guidelines suggested that MCQs should be included as one paper for PGs - in 2004 - this has been excluded
- MCQs at PG can be useful if the quality of MCQ is good - Currently quality of MCQs in Indian exams is far from satisfactory - hence most universities do not use MCQs for PGs. Our own university has deferred decision
- It really needs lot of effort and time from teachers to create new MCQs for PGs - we all find this difficult. What is the experience of others?

✉**Vivek Saoji:** To answer Venu's specific query on MCQs: Yes, MCQs can and should be used in summative exams both at UG and PG level and not confined to some formative assessment. It has a lot of advantages and most of them have been mentioned in the discussion so far, like they are objectively marked, they sample the content adequately, they are easy to score, store, reuse and so on, so not only they have good psychometric properties but they assess examinee comprehensively and this is all backed by more then 50 years of research.

So why don't we change over, unfortunately the quality of MCQs we produce are almost all flawed, also they test only the lowest cognitive domain i.e. recall type and occasionally comprehension, there are flaws like unfocused stem, implausible distracters, use of all of the above, none of the above, negative stem etc.thus lowering the quality of MCQs.

The bottom line is we have to produce good quality MCQs. Literature is abundant with item writing guidelines and Payal has mentioned the excellent resource, there are also n number of studies about item writing flaws and its impact on quality of assessment and validity. So rather then criticizing MCQs we must make a positive efforts to improve its quality based on scientific guidelines and again the bottom line here is urgent need for faculty development in item writing.

Finally there was a comment in the discussion about, if it is all that difficult is it worth the efforts, again my answer to this is an emphatic yes, if we want to improve the way in which we want to assess our students then its definitely worth and to quote from the bible on medical education "Good selected - response items are difficult to create well and are not as time efficient to write, as are some of the constructed - response formats, such as essays. However, time and cost efficiencies lost in the test development process are more then adequately compensated for in the scoring phase of the testing cycle" (from International Handbook of Research in Medical Education pp.653).

I would conclude by saying that, MCQ is a definitely better assessment tool for the assessment of cognitive domain but it is not a panacea and a combination of methods is always good

✉**Vivek Saoji:** Yes, preparing good quality MCQs is difficult and time consuming as I wrote in my previous communication, but then we spend enough time in preparing for our lectures, seminars, tutorials, or teaching activities in general. Do we spend at least 10% of that time in planning or assessing our students. I am sure if we give little more time and thought for student assessment, take the assessment more seriously then it will go a long way in overall improving the teaching - learning process. It is definitely possible but we have to give more time to it.

✉**Monika Sharma: H**ere is a copy of an article from the Website of Oregon University that gives hints on how to prepare MCQs that test the critical thinking of the student.

✉ **Chetna Desai**: Here is an article on MCQs with focus on item analysis.

✉**Sanjay Bedi**: Sending a nice presentation on MCQ Test Construction.

OSCE

✉**Chetna Desai:** OCSE does have advantages in terms of objectivity, uniformity of assessment and discipline with time and procedure that is maintained. However it is also felt by many that it makes the examination a bit robotic and narrow in context. Since it takes a lot of time to prepare and implement, a wide range of topics cannot be covered. Also it is time consuming and may not be suitable for large batches. Of course once the examiners are familiar with the system and the stations, it flows smoothly.

✉**Payal Bansal**: I think OSCE allows for wider sampling, and, like MCQs, if well constructed can measure complex learning. It is well accepted and widely used the world over for clinical skill assessment. Of course, the assessment can be a composite one, with OSCE being one of the components.
✉**Stewart Mennin:** Perhaps multiple methods can strengthen the assessment program. The OSCE is time consuming and it offers, when there are enough stations, a picture of student ability not accessible in most other ways, unless you are developing a miniCEX also.

✉ **Tejinder Singh:** Like MCQs, the answer lies in devoting enough time in constructing and refining the stations. Otherwise, it becomes no better than spotting and as Ara once said, looks like an artefact than an OSCE.

✉**Tejinder Singh:** I am going to contradict some of the issues raised by Chetna. OSCE in fact is better suited for large batches. Imagine trying to take a conventional case

presentation for 150 students. The minimum you will need is 7 days non stop. When I was incharge of postgraduate selection at CMCL, we introduced OSCE in the selection procedure. There were 15 stations and in some years, we had over 300 students taking the test in one day. However, you are right that it takes a lot of time and effort to design, co-ordinate, administer and evaluate the performance.

✉ **Monika Sharma**: I agree with Payal on this one. I have been given the responsibility of conduction of sendups of students in my department, and since OSCEs are the most labor intensive, everybody else is quick to decide on taking up theories or practical co-ordinations, while I end up being the one to conduct OSCEs. With each one conducted, I find it easier to prepare and conduct,.not because I think I have mastered the art, but because I have started understanding the purpose of it and the language required to prepare it.
Just like MCQs, preparing an OSCE station means you need to be specific about the objective of each. Besides isn't it better than MCQs? While MCQs evaluate the knowledge mainly, OSCEs can be modified to evaluate psychomotor skills as well as affective domain, can be used for UGs as well as PGs (is an essential and actually the more difficult part of the national board examinations).

Questionnaires

✉**Anshu:** Can you tell me about situations where questionnaires work best as assessment tools? Also, what can be done to improve the reliability and validity of responses obtained from these questionnaires?

✉**Monika:** I plan to use questionnaires and feedbacks in my project to evaluate its feasibility, usability etc. It is valid as it would give the true impressions of the users, but reliability is suspect. The responses of the observers may vary upon several things- my influence on the students, my relation with them, their attitude towards me(a student who has been at the receiving end of hard words from me, might try to get even by giving responses that he feels may ‘spoil’ my project.). Well there could be some more factors.

✉**Tejinder Singh**: While there are many references available on this, I am referring you to our paper cited below-

- M. Verma, Tejinder Singh
  Designing attitude scales: Part -I - Theoretical considerations .Indian Pediatrics November 1993; 30 (11): 1369

- M. Verma, V. Saini, Tejinder Singh
  Designing attitude scales: Part II - Developing and standardizing. Indian Pediatrics November 1993; 30 (11): 1303

I hope, you will find them useful.

✉**Payal Bansal:** Regarding reliability and validity of questionnaires: Reliability is calculated the same way as for an assessment. Improving validity - means spending more time and effort in designing the questionnaire in a systematic way - as for MCQs, both content (of questions) and process(wording/framing the question and responses appropriately) are important in questionnaire design as well.

Mini CEX

✉**Tejinder Singh:** Mini CEX stands for mini clinical examination and is a tool designed for assessing a snapshot of professional behaviors by the bedside. Make a google search and you will find plenty of references. Mini CEX can be a good point for discussion, specially when a number of institutions have reported encouraging results with its use.

✉**Avinash Supe:** Minicex- is clinical attitude testing method - tests attitudes and behaviours through 15 minutes - doctor and patient interaction.

Sending you link of Dr Norcinis article - very well cited –

✉**Monika Sharma:** It was an excellent article on mini-CEX, and an encouraging new concept.

1. It does imply that mini CEX can be used for evaluation of post graduates.

2. Most of us are not convinced about the reliability and validity of the conventional 'long cases' as a single contact evaluation of the learning done over a period of 3 years, as several factors at the time of the final exam may benefit or be a 'big unlucky day' for some. It could be part of an internal assessment, though the concept of internal evaluation for PGs is not existent formally at the present. Factually speaking, personal/academic influences of residents created on the 'internal examiners' does act as an influence on the performance and the final results in the long case presentation. Rather than allowing this practice of 'impressions', mini-CEX can be used as an objective way of internal assessment. It may be incorporated in the form of ratings of residents made available to the examiners prior to the final exam. I really don't know if this can be a viable option

3. Secondly, as we all know it may take ages before somebody in MCI is impressed to actually implement it. Until then mini-CEX can be used as eye openers too. A trial mini-CEX on all the residents in a department once or twice a year can tell us where our residents are lacking or concentrating. For example, it may reveal that some residents are doing too much of talking and counseling (wrongly many times) rather than actually examining patients and trying to come to the right diagnosis, or at other times reveal that PGs are

trying to come to a commoner diagnosis and treating it fast, rather than spending time explaining to the patients.

4. Everything has limitations. To implement mini-CEX, like any other evaluation tool, faculty needs to be prepared for the technique, devote time and attention. To evaluate residents in different scenarios we need to move around a little more to observe the residents, in OPDs, emergencies or intensive care. Obviously it cannot be a part of summative evaluation. As of now, our PG programmes have no place for formative evaluations like the western set up and we solely rely on the single final exam. Like the OMP I would like to call it the OCP-one case 'per'ceptor...you see hundreds of cases in three years, you read hundreds of pages in three years, you present several cases to various faculty members just to practice, and when it comes to pass-fail, it is the one case through which the examiners 'per'ceive your level of knowledge, competence and whatever!

5. I would consider miniCEX as one way to enlighten the department of its teaching-learning fallacies/deficiencies.

6. MiniCEX can be performed during the daily rounds alike bedside teaching. How is the idea of one miniCEX every morning? It would be a faster means of showing the PGs what and where they lack.

7. If the students know, how they are being rated, would it positively or negatively influence their performance and manipulate the observer?

I found the idea impressive, practical and practically usable in the midst of so many outdated, rusted clinical evaluation methods.

✉**Tejinder Singh:** Good to see one more convert(!) However, like any other tool, mini CEX requires preparing the faculty as well as the student to give and receive feedback. One or two such encounters every six months may be enough.

✉**Stewart Mennin:** The miniCEX was originally used for formative assessment. I don't think it is meant to be used every morning. It would be best used between 1-4 times per most graduate in each rotation or area. Then, it would be good to use it in several areas, internal medicine, pediatrics, emergency, etc. Óne consideration is to do a pilot introduction of the miniCEX with a small group of post grads and teachers and then stimulate interest among others, spread the word etc., a mini campaign.

✉ **Avinash Supe:** As Stewart has said - MiniCEX has to be done only few times in PG curriculum - at intervals of few months

Web Based Evaluation

✉**Chetna Desai:** Web based evaluation is an important topic that can be discussed. What are its advantages, disadvantages and feasibility. In India too online examinations are gradually being introduced. Apart from infrastructure and computer literacy, what are the other factors that act as a deterrent to this assessment tool?

✉**Payal Bansal:** The advantage would be ease of administration, efficiency and convenience to examinees. To me, in our set up, exam security seems to be an important deterrent.

✉**Tejinder Singh:** Though I am using computers a lot, I am a bit hesitant in using them for any and everything for some strong reasons. Web based or online examinations are only a delivery mode- the basic product still remains the quality of assessment tools. A poor quality MCQ will remain a poor quality whether used on paper or over the web. Technology can not compensate for poor quality.

I feel, that first we should ensure faculty development to make sure that our teachers know how to write good items/questions and then only should we venture into online examinations.

Regarding security issues, probably they will be little more secure than paper tests but this consideration is only secondary to quality issue.

Portfolio

✉ **Payal Bansal**: As suggested by Dr. Tejinder, I am attaching a chapter on portfolios that I have written as part of a compliation of readings being prepared for FAIMER Regional Institute Fellows.

✉ **Rita Sood**: Portfolio assessment (collection of evidence that learning has taken place) though labour intensive is a method gaining popularity worldwide for assessment of qualities which are difficult to assess using most other methods like the skills of reflection, critical thinking, professionalism and attitudes.

Lots of literature is available on portfolio assessments in medical education. Margery Davis from Dundee Institute has pioneered the use of portfolios and conducts workshops on the use of portfolios.

✉**Sheena Singh**: I have read the excellent article on Portfolios. I recall that as a student many of us used to share our reflections in informal chats however we did not document them or utilize them as an assessment tool. I agree that some people are more reflective and analytical; it must be a challenge to tutor people to do this.

We are all familiar with the exercise of writing our Career portfolios or CVs. There must be formats for writing Portfolios. It would be nice to see an example of one. I suggest we all write our own FAIMER Portfolios as we continue to journey forward.

✉**Payal Bansal:** I am posting a sample reflection, which gives an idea of how a reflective piece is written (and evaluated).

Standardized patients

✉**Payal Bansal:** To complete the toolbox checklist, an important "miss" has been standardized patients. We all seem to think that since we in India do not have a dearth of patients, we don't need standardized patients. However, one of the greatest challenges to validity and reliability of our existing clinical exams is non-uniformity of the clinical exam as each student gets a different case. This is where standardized patients come in. Elsewhere, e.g. in USMLE, which is a high stakes exam, the clinical skills assessment is largely based on standardized patients.

✉**Monika Sharma**: The concept of standardised patients sounds useful, specially because it wd remove the disparity in performance of students on account of the patients behaviour and other patient related factors. I think we do use a touch of impersonation in some of the OSCEs we conduct, like asking a nurse or a resident to sit at a station and act like a parent with standardized answers to certain questions likely to be put by the students. Since this is a small scale attempts it seems to be fairly reliable, doing the same for the long case set up. I am not sure if we have the required training or can give the right kind of training to the standard patients. I have no exposure to USMLE though I have read about the same. The concept is encouraging but what about practical problems?

✉**Payal Bansal:** Making an exam uniform is a basic criteria for effective and meaningful assessments. If not standardized patients, then what? Can we think of an alternative? It is difficult.

So why not identify situations where we can use standardized patients and still be close to real life?

For example, we have used a standardized patient (SP) to evaluate history taking skills in a patient with sudden onset lower abdominal pain in one of our exams.
Now, in our regular exams, we only have "cold" cases, so we never end up measuring this ability, which is perhaps the most commonly encountered surgical situation.

We developed a case scenario, trained student volunteers (once medical students, another time nursing students) for role portrayal for acute abdominal pain, and then conducted the OSCE, of which this was one station.The station showed good reliability.

Another example: To test if residents/interns could paint and drape a patient for a basilic cutdown without break in aseptic technique, we had our ward-boys volunteer as SP's.
It is all doable - only, as Venu has rightly said, it should be relevant and meaningful for our situation.

Can you think of other situations in your speciality where SP's can be used? Is there any other way to make clinical/performance assessment uniform?

✉**Monika Sharma**: We are using some already in OSCEs. To cite a few examples,

- Our office clerk was used as a distressed mother of a child suffering from a malignancy at an OSCE station for counseling.
- Nurses after a short explanation act as mothers of vaccinated children at OSCE stations where students are asked to explain the side effects and advice for a common vaccine like DPT.
- Mothers of babies in our ward help as SPs at OSCE stations where history taking skills are being tested.
- Normal patients can also act as SPs for testing certain clinical skills like how to take blood pressure or anthropometric data.

Marks Vs Grading System in evaluation

✉**Venugopal Rao:** I have been as PG examiner for the past 8 years. Previously we used to have grading system in PG degree Evaluation in Final exams. For the last 5 yrs we are using marks system. I feel somehow it is not the correct method. The difference between good and average is minimal, and with 1 or 2 marks students are losing, especially in theory.
How far this is reliable and valid? How about grading system?

✉**Mrunal Ketkar:** I am originally from Mumbai University. When I appeared for my P.G. we had something called as the point system.6 points for theory and another 6 for practicals. It was kind of vague. No explanations,no grading and it especially was difficult to apply for a job as there was no marklist available or anything to support the points you were claiming to have got. I lost my initial job as I didn't convert my points. We seem to have different evaluation systems all over India. How far are they comparable? Which is the best?

✉**Dinesh Badyal:** I know about PGI, Chandigarh. We used to get a percentage for final university examination, but this was not just based on our performance on the day of examination.
For every activity by a PG in department i.e. seminar, journal club, monthly theory test, weekly practical etc. a percentage was awarded. This percentage was based on evaluation form filled by all faculty & PGs who were present during that activity. So the overall percentage was based on what you have done in 3 years of residency. All these activities were very regular, so even a faculty members resigns/joins it did not affect the system. That was in fact good- bias was removed to an extent.

Nowadays I'm working in Punjab. Here we have marks 400 for theory and 400 for practical (100 viva +300 practical conduct). I find this system quite faulty, because there are no regular activities, so a teacher depending on his relationship with a particular PG gives marks. I think atleast grading will be better. But till PGs are given proper training, and are not considered personal property, and if not evaluated regularly at short intervals- whatever system we adopt at the end result will be same i.e. HOD's happiness with PG.

✉**Madanlal Gill:** I know of some institutes which offer grades for evaluation because they bring healthy competition rather than negative comparison.

Self assessment tools

✉**Chetna Desai**: Self-assessment devices are likely to be particularly valuable educational tools. These have been used in a variety of disciplines as professional development aids. Formative uses of self-assessment focus on individual learning, particularly to reinforce behavior change. These tools are typically designed to accomplish the following aims:

- allow users to reflect on their own performance strengths and weaknesses to identify learning needs
- reinforce new skills or behaviors to improve performance.

- However self assessment tools are have certain drawbacks since a perception of improvement by the user may not reflect the actual improvement when tested by other tests. Hence how good are these tools and when can they be used? Any answers??

Meanwhile sharing two articles that describe both sides of the coin....
http://www.bmj.com/cgi/content/full/315/7120/1426
http://www.sanswired.com/?&function=change_panel&new_panel=about

✉**Chetna Desai:** I believe that students also need to assess themselves periodically since the medical curriculum does not allow for frequent evaluation. We for example have three internal exams and three system ending tests during 18 months of pharmacology- really not adequate. Moreover our students are encouraged to adopt certain adult learning principles, then why not self assessment too? Interested and motivated students should be allowed some self assessment.

✉**Payal Bansal:** Reflection is a good way of self assessment. See the reflective piece I have attached with "portfolios"
✉**Sanjay Bedi:** I was participating in a US based University Intranet discussion and came across this form which is filled by Postdoctoral fellows and in fact everybody. Is there any such practice being followed in India anywhere?

EVALUATION OF FACULTY RESEARCH TEACHING

FACULTY NAME: _____

FELLOW/STUDENT NAME: _____

DATE: _____

STATUS WHILE IN PROGRAM (check):

GRADUATE STUDENT _____ *POSTDOCTORAL FELLOW* _____

MEDICAL STUDENT _____ *UNDERGRADUATE STUDENT* _____

Please answer the following questions in relation to the training provided by your laboratory mentor and the program while in the Department of Pathology using the rating scale below. Write your ratings in the blank provided in the left margin.

1 2 3 4 5 N/A

Poor Adequate Outstanding Not Applicable

Rating

_____*1. Your mentor's knowledge of the subject area.*

_____*2. Level of intellectual stimulation provided by your mentor.*

_____*3. Level of the intellectual atmosphere within your mentor's laboratory.*

_____*4. Quality of the intellectual atmosphere within the Program (including faculty, students, fellows).*

_____*5. Quality of the instruction provided by your mentor in developing a logical approach to scientific investigation.*

_____*6. Quality of the feedback and criticisms provided by your mentor.*

Rating

_____*7. Quality of the scientific activity you were engaged in and methodologies (state-of-the-art) employed.*

_____*8. Quality of the physical resources available to you to perform your laboratory studies.*

9. How well did your training prepare your for further development as an independent investigator and/or fulfillment of your career goals? (Use the same 5-point scale as above.)

_____*A. Scientific Process*

_____*B. Knowledge of Subject Area*

_____*C. Writing Skills*

_____*D. Oral Presentation Skills*

_____*E. Grant Writing*

_____*F. Teaching Skills*

_____*G. Career Counseling*

_____*10. What was the overall quality of training and direction provided in scientific investigation by your mentor?*

_____*11. How would you rate your overall experience within the Training Program/Lab in Pathology?*

Please include any additional comments related to your training while in the Department of Pathology.

(Please forward the evaluations directly to Student Services Assistant, , in the Pathology Education Office, who will collect the evaluations and assure confidentiality.)

✉**Stewart Mennin:** The structure of questionnaires about the experience of a course or event is variable. The way the questions are phrased is important. I prefer to make statements and then the respondent can indicate 1 disagree very much, 2 disagree, 3 neither disagree nor agree, 4 agree, 5 agree very much. For example
___ *My mentor's knowledge of the subject area was accessible to me*

___*The Level of intellectual stimulation provided by my mentor helped me to learn*

etc. In this way you know what the answer means. Interpreting adequate or poor is variable and one person's adequate is another person's poor. Knowing the extent of agreement or disagreement with a statement, particularly one focused on the desired outcomes, lends itself to providing a focused discussion and formative feedback.

✉**Dinesh Badyal:** Most of times we evaluate students to see success of intervention. A few institutes are doing this. It is not exactly validity, but there is an effort to see what a teacher is doing, it is mix of reliability and validity. The management at the Himalayan Institute at Dehradun, is evaluating faculty for promotion. Students evaluate all teachers after passing a particular professional exam. Similarly in DMC Ludhiana, student feedback is taken for all teachers as well as department, from passed out students for a particular year. A questionnaire is used. All teachers and department are informed of their performance. There is some resistance from faculty. But the institute does it regularly.It shows beneficial effects.

✉**Tejinder Singh:** We can assess performance of the teachers but not validity of the teaching. There are a number of factors which come into play in student ratings. One of the teachers at CMC, for example was using lot of notes and questions to prepare students to write well in the examinations. The topics covered were mostly 'important' ones, which commonly are asked in the examination. However, a number of other must know areas were left out. This teacher was very popular with the students. Can we say that his teaching was valid?
To me, the best way of infering validity of teaching will be to correlate it with learning objectives and then see if all the objectives were taken care of by appropriate methods. However, like validity in general, it is difficult to numerically express it.

✉ **Hemlata Badyal:** In our institute, students after passing every professional give a feedback on faculty and departments. It is based on a questionnaire. It is done by Principal's office. The individual faculty members get a confidential letter telling him about his/her position in basic/paraclinical/clinical faculty, a percentage is given. Teachers also get photocopies of best/worst comments of students along with that letter.

Each dept HOD gets a letter telling him/her about dept position. Initially there was resistance from faculty, other problems also surfaced like some teachers started buttering students to improve their ranking. But as you know students are quite good in picking these. Ultimately only good teachers are getting good ranking- all other factors now stand diluted. But it is now the 5th year of student feedback process, a continuous effort from the Principal's office. The beneficial effects which are now visible are that every teacher is now improving areas highlighted by students. There is no other way.

✉**Sheena Singh:** The format for the Teachers Evaluation by Residents was interesting. It reminded me of a PG Seminar Evaluation and Feedback form that I had made a year ago.I am attaching it for you to look at

✉**Sanjay Bedi**: A last minute addition of two nice questionnaires I found on the site http://www.hkca.edu.hk/ita.htm

The Final Word

✉**Balchandra Adkoli:** As an educationist rooted in the philosophy of "Constructivism", I hold the view that people construct their own "truth" of the reality. No one knows what is absolute reality. We try to negotiate with reality through our own perception, experience and intuition. The debates regarding different forms of validity and reliability will continue for ever. Even if there is an element of more or less confusion at the end of the day, we are winners, because we made an effort to understand.

Take Home Message

- Choosing a feasible assessment tool is one of the most difficult decisions to make. The success of any assessment depends on finding a suitable balance between validity and reliability and between the ideal and the practical.

- Validity is never assumed. It uses theory, logic and scientific methods to collect and assemble data to support or fail to support proposed score interpretations at a given point in time. Though not always practical, one must gather as much relevant evidence as is feasible with the constraints of the situation. We should also look for various types of evidence when evaluating standardized tests:

- A reliable test result is a prerequisite to a valid test result. One can have a reliable test score that is not valid, but one cannot have a valid test result that is not reliable.

- Whether an assessment tool is valid or not depends on the purpose for which it is used.

- It is always a challenge to identify the right methods to assess the different levels of competence, which are not only able to assess the examinees in all the domains but have good psychometric properties so to say that they are valid and reliable.

- Assessment deserves as much attention, patience and thought as any other research.

Suggested Reading:

1. Principles of Medical Education (Dr Tejinder Singh et al):
2. Downing's articles on validity and reliability
3. Toolbox of Assessment Methods

PARTICIPANTS

1. Anshu
2. Aroma Oberoi
3. Avinash Supe
4. Balachandra Adkoli
5. Chandrika Rao
6. Chetna Desai
7. Dinesh Badyal
8. Hemlata Badyal
9. Jayanthi V
10. Madanlal Gill
11. Monika Sharma
12. Mrunal Ketkar
13. Nirmala Rege
14. Payal Bansal
15. Rita Sood
16. Sanjay Bedi
17. Sheena Singh
18. Stewart Mennin
19. Tejinder Singh
20. Varughese P Varkey
21. Venugopal Rao
22. Vivek Saoji
23. William Burdick

Congratulations on a fantastic discussion of the concepts of reliability and validity! I think the CMCL-FAIMER Regional Institute 2006/2007 Fellows and their faculty should get a dozen gold stars for a stimulating, provocative on-line interaction.

Keep up the good work on the listserv. Your postings have been excellent in the information they have provided and the questions they have asked. Your discussion can serve as an exemplar for the rest of us!

William P. Burdick, M.D., M.S.Ed.
Associate Vice President for Education
Foundation for Advancement of International Medical Education and Research