Van Blerkom Chapter 2: Criterion- and Norm-Referenced Interpretations

1. Frames of Reference

To interpret performance, one needs a reference. The following four references are often used in education.

a. Ability-referenced

* This frame of reference is based upon an individual’s ability or skill level in a given area; performance is compared against what one judges a student should be able to do given their ability.
* Ability-referenced interpretations are usually fruitless because it is often very difficult to assess, accurately, students' skills and abilities.
* We will learn later that ability-referenced grading should be avoided.

b. Growth-referenced

* If one is interested in academic or performance change or improvement from prior levels, then one is referring to growth-reference interpretations.
* From a statistical point of view, growth-referenced interpretations may be problematic because to measure growth, at least two tests or assessments are needed, and both assessments will have measurement error, so the difference score – growth sore – calculated between the first and second assessment will have larger measurement error thus making it less precise.
* Another potential problem is the learning curve which can differ among students – some learn faster or slower than others. This difference in learning curves also reduces the utility of growth-reference interpretation growth scores can mea different things. For example, if two students obtain the same score on the final test, say a score of 100% correct, but one student start the pre-test with a score of 70% and the other a score of 50%, then the second student showed more growth, 100 – 50 = 50 vs 100 – 70 = 30. This difference in growth ignores the fact both achieved perfect scores and instead if growth scores are the focus the second student is judged superior.

In general, ability- and growth-reference interpretations should be used only for formative evaluations, and are typically best applied in one-to-one situations wth students rather than to whole groups of students.

c. Norm-referenced

Norm-referenced interpretations answer questions such as

* “how well did I do compared to the class”
* “did you do as well as someone else”

Norm-referenced interpretations enable one to assess a student’s performance relative to some comparison group, so, normed interpretations enable one to determine

(a) what is typical, common, or average, and

(b) what should be expected or what is reasonable relative to a comparison group or norm group.

A primary limitation of norm-referenced interpretations is that such scores do not indicate what students have learned or mastered. A percentile rank of 100, for example, does not tell us whether those who achieve this rank understand all, most, or some of the content tested, instead, it only tells us that among the normed group, these students scored equal to or greater than all others.

d. Criterion-referenced

In education, and specifically in classrooms, criterion-referenced interpretations are based on student performance relative to preset standard or criteria for some **content domain**, such as identifying appropriate sampling procedures for a given research design, or determining the appropriate forms of reliability given data source used.

Such performance is oftentimes set by well-defined domains, for example

* the student can, with 90% accuracy, correctly determine the appropriate sampling strategy when given 50 research scenarios;
* the student correctly determines all which form of reliability to be used given various measurement problems

Thus, criterion-referenced means that one, such as the instructor, directly describes the specific performance that is to be demonstrated (for example, type 30 words per minute without error), so the criterion-referenced interpretation enables one to describe what an individual can do without reference to others’ performances.

In short, criterion-referenced interpretations enable one to know exactly what one can and cannot do relative to some specified content domain. With criterion-referenced scoring, students know how well they are doing on content. A potential problem with criterion-referenced tests is the difficulty of developing tests that meet the standards to be considered criterion-referenced (i.e., specified content domain, objectives, and adequate number of item to sample each objective).

Examples: identify whether normed or criterion-referenced:

* you are my favorite grandchild (there are many grandchildren) - normed
* correctly counts from 1 to 100 - criterion
* identified 15 of the 23 medical instruments - criterion
* Jim’s IQ is at the 63rd percentile - normed

Note that criterion-reference interpretations require that one explain those specific skills measured, that one clearly identify the content objectives that must be met. For this reason, the following cannot be considered a criterion-referenced interpretation:

Joe made an 83 on her statistics test.

2. Appropriate Interpretations for Evaluations

a. Diagnostic Evaluations

Since these types of evaluations are important for determining specific weaknesses of students, the goal, then, is to learn what students can and cannot do--and this is most usually criterion-related. Note that this requires that skills to be assessed with diagnostic tools that must be well-defined; it is also possible for diagnostic tools to indicate whether a student’s problem is common or typical, and this is norm-referenced.

b. Formative Evaluations

Recall that formative evaluations are most often used to determine appropriate instruction action or alteration, and this is based upon what the student has or has not achieved; therefore, formative evaluations require criterion-related interpretations.

c. Summative Evaluations

Because summative evaluations use global tests (midterm or final examinations), and because these types of tests cover many domains with only a small sample of items, and because of this limited item sampling from many domains, it is difficult to clearly articulate what the student can and cannot do, so norm-related interpretations are usually most suitable. However, the more tests and other assessment activities collected, and the better defined the domains covered by each measurement, the more likely one can use criterion-referenced interpretations for summative evaluations.

3. Qualities for Norm- and Criterion-referenced Classroom Tests

Much detail will be covered on these qualities later in this course, however, three can be covered now:

a. Item difficulty

In short, easy items (items in which most all students answer correctly) provide little information in terms of what the student does not know, so easy items typically should not be included in either normed (because easy item inflate scores and make it more difficult to establish true norms) or criterion tests (because easy items do not indicate what students do not know – an important goal of criterion tests).

For similar reasons, extremely difficult items should also be avoided. Moderately difficult items are good for criterion tests, and slightly more difficult items are good for normed tests.

b. Sensitivity to Instruction

Sensitivity to instruction refers to whether a correct response to an item requires some type of preceding instruction. The issue of whether items should be sensitive to instruction does not revolve around the normed vs. criterion issue; rather, sensitivity depends upon the nature of the skills being assessed. For example, it is very easy to develop items that are highly sensitive to instruction – simple recall of facts type items match this description well; however, should you be more interested in testing high-order intellectual skills, it is unlikely that such items will be highly sensitive to instruction as higher-order skills are often not taught in typical classrooms.

c. Number of Items on a Test

Since criterion-related tests require well-define domains, and since making statements about students' performance in such domains requires adequate sampling from such a domain, the number of items for criterion-referenced tests needs to be large. For norm-referenced tests, fewer items are needed since the goal is to determine group-relative performance, not statements about a student's performance within a given domain.

Self-Test

1. How do norm- and criterion-referenced interpretations differ?

2. For each of the following, indicate whether a norm-referenced of criterion-referenced interpretation is preferable:

1. Compare the performance of students in your school district to the performance of a nationwide sample of students in basic skills.
2. Plan lessons covering the punctuation of declarative, interrogative, exclamatory, and imperative sentences.
3. Compare a student's performance across reading, language, and mathematics skills.
4. Analyze the variability of students' performance in mathematics within and across grades.
5. Diagnose problems a student or class has in performing district-prescribed instructional goals.
6. Determine students' mastery of selected language usage skills.

3. Explain how the following three items relate to criterion-referenced tests:

1. item difficulty
2. sensitivity to instruction
3. number of items.