The Table of Specifications: Insuring Accountability in Teacher Made Tests

Charles E Notar, Dennis C. Zuelke and Janell D. Wilson and Barbara D. Yunker

Teachers have been in the era of accountability for some time. There is an increased demand for accountability and the use of non-referenced testing with President Bush's "No Child Left Behind" initiatives. However, there is a growing demand for less reliance on standardized tests. Admission decisions to colleges and universities are being made with less emphasis on using standardized test scores and more on other criteria such as Grade Point Averages (GPAs). GPA is a standard of accountability. However, when you compare GPA and standardized test scores there are frequently differences among students GPA and scores on a standardized test, sometimes very large differences. From the literature we know standardized tests are valid. The question needs to he asked if GPAs are a valid measures of student achievement. GPAs are based in large measure on teacher made tests. If teacher made tests are not valid, how can a students GPA be valid? This paper looks at teacher made tests and validity. The use of a Table of Specifications can provide teacher made tests validity. This paper provides the why a Table should be used and how to construct a Table for their assessment purposes.

The literature is full of articles on accountability issues in education (Eisenberg & Serim, 2002). Others agree. Mehrens and Lehman referred to the "age of accountability..." as far back as 1973. More recently, Falk (2002) Nathan (2000) and Newell (2002) have spoken to the growing demand for accountability given the massive use of norm referenced testing in today's schools. The literature is full of articles on norm-referenced achievement testing. The literature is full of articles on admission policies and the selection decisions being made on the basis of standardized test scores and grade point averages (Imber, 2002; Jenkins, 1992; Marshall, 1997; Micceri, 2001; Patton, 1998; and Perfetto, 2002).

However, the literature is not full of accountability issues regarding teacher made classroom tests. Teacher made tests have flirted with, had affairs with and been engaged to accountability, but a permanent relationship has not materialized. Now, it is the time for a marriage to take place. The reasoning is simple - the GPA. The grade point average (GPA) is a standard of accountability based on course grades resulting from teacher made, or teacher chosen, content specific tests. And, although the GPA may be considered in selection processes, norm referenced test results may carry more weight. This happens because there is too often a poor relationship between the GPA and scores on norm-referenced achievement tests.

Lei, Bassiri and Schultz,(2001) found that a college GPA was an unreliable predictor of student achievement. Since we assume that norm referenced tests are valid measures, the tendency is to put more weight on those results concerning student achievement. Opponents of standardized achievement testing would argue otherwise. For example, Bennett, Wesley and Dana-Wesley (1999) suggested

Charles E Notar, Ed D., Assistant Professor, Secondary Education. Dennis C. Zuelke, Ph. D., Professor, Educational Administration Janell D. Wilson, Ph. D., Associate Professor, Secondary Education. Barbara D. Yunker, Ph.D., Associate Professor, Secondary Education, Jacksonville State University.

Correspondence concerning this article should be addressed to Dr. Janell D. Wilson, Ph.D., Associate Professor, Secondary Education; Email: jwilson@jsucc.jsu.edu

that a college admission model should be developed to encompass GPA, rank in class and a district performance index or a similar predictor as an alternative to standardized test scores. A formula index based on these predictors would afford some protection in selectivity issues. But, since a GPA may not significantly correlate with norm referenced test results, which measure is the more valid measure? The belief in the validity of norm referenced achievement tests today is strong. However, can we tell if the GPA is a valid measure? Where is the validity data for the teacher made (or chosen) tests on which GPA is calculated? If the teacher made/chosen test is NOT valid, the GPA will not be valid either.

Therefore a marriage between teacher made tests and accountability should take place to insure validity of its offspring, the GPA. For a wedding you need something old, something new, something borrowed and something blue. The blue is relatively easy to find. It is the blue of the student who may be taking a teacher made test that lacks validity. Teachers who do not use conventional construction guidelines for paper/pencil test development will not be assessing student achievement well. Their tests will likely have poor content validity, "cause for concern because each assessment instrument depends on its validity more than on any other factor." (Ooster, 2003, p. 40)

The something borrowed is a Table of Specifications. In assessment literature, the Table may also be referred to as the "test blueprint," "master chart," "matrix of content and behaviors," "prescription," "recipe," "road map," "test specifications," or "formal specifications" (Bloom, Hastings, & Madaus, 1971; Carey, 1988; Gredler, 1999; Kubiszyn & Borich, 2003; Linn & Grunland, 2000; Mehrens & Lehman, 1973. Ooster, 2003). We prefer the concept of a test blueprint.

A blueprint is a crucial concept when constructing anything. First, it is important to know what you are building before you start. Constructing an outbuilding for a riding lawnmower and building a house are very different projects with the first requiring something more than a sketch; the second requiring a full set of detailed plans. In testing terms, a quiz does not need as much detailed attention as a unit or a grading period exam.

The Table is being borrowed because norm-referenced achievement tests are usually constructed from such a blueprint. The blueprint is meant to insure content validity. Content validity is the most important factor in constructing an achievement test. The most important factor in determining the GPA is the teacher made achievement test called the end of unit test. A unit test or comprehensive exam is based on several lessons and/or chapters in a book supposedly reflecting a balance between content areas and learning levels (objectives). The Table serves to clearly define the scope and the focus of the test. The Table insures correspondence between the learning objectives for the students and the content of the course. A Table serves to organize the process of test development to best represent the material covered in the teaching/learning process. Without a Table or a test blueprint, a test will produce scores of limited use and interpretation.

A Table of Specifications consists of a two-way chart or grid (Kubiszyn & Borich, 2003; Linn & Gronlund, 2000; Mehrens & Lehman, 1973; Ooster, 2003) relating instructional objectives to the instructional content. The column of the chart lists the objectives or "levels of skills" (Gredler, 1999, p.268) to be addressed; the rows list the key concepts or content the test is to measure. According to Bloom, et al. (1971), "We have found it useful to represent the relation of content and behaviors in the form of a two dimensional table with the objectives on one axis, the content on the other. The cells in the table then represent the specific content in relation to a particular objective or behavior" (p.14).

Teachers often use performance objectives to guide instruction and subsequent test item construction. However, this tactic too often results in test items measuring rote

memory only. In order to measure students' achievement at the higher learning levels of comprehension, application, analysis, synthesis and evaluation, teachers should go one step further. Teachers should make use of the test blueprint - the Table of Specifications. A Table of Specifications identifies not only the content areas covered in class, it identifies the performance objectives at each level of the cognitive domain of Bloom's Taxonomy. Teachers can be assured that they are measuring students' learning across a wide range of content and readings as well as cognitive processes requiring higher order thinking, The use of a Table insures that teachers include test items that tap different levels of cognitive complexity when measuring students' achievement. Kubiszyn & Borich (2003) suggested that teachers should use a Table so they won't forget the details.

Carey (1988) listed six major elements that should be attended to in developing a Table of Specifications for a comprehensive end of unit exam: (1) balance among the goals selected for the exam; (2) balance among the levels of learning; (3) the test format; (4) the total number of items; (5) the number of test items for each goal and level of learning; and (6) the enabling skills to be selected from each goal framework. A Table of Specifications incorporating these six elements will result in a "comprehensive posttest that represents each unit and is balanced by goals and levels of learning" (p. 89).

A Table of Specifications is developed before the test is written. In fact it should be constructed before the actual teaching begins (Kubiszyn & Borich, 2003; Mehrans& Lehman, 1973; Ooster, 2003). As much time and effort is spent on developing the house blueprint; so too a Table of Specifications requires considerable time and effort to develop (Kubiszyn & Borich, 2003). Linn and Gronlund (2000) stated "While the process is time-consuming, the effort that goes into development of a table of specifications also makes it much easier to prepare the test once the plan is developed" (p. 147).

Table 1 is an example of a basic table of specifications.

Heading provides the administrative data for the test and Table. All tables of specifications have a Table heading. The heading provides for the administrative requirements of the test and the information needed to construct the two-way table. The heading makes it easier for filing and retrieving.

The course title is exactly that, the title of the course as seen on the teachers' and students' schedule e.g. American history I, English II. Grade level is the grade for which the course is intended on the local or state course of study. Test periods are time limits for which the test has been developed for administration. Date of test is the date the teacher will administer the test.

The subject matter digest is a paragraph that provides the limits of the subject matter that will be covered in class. This insures that the class covers only required material as related to stated objectives and nothing else. This setting of parameters helps guide discussion and keeps lessons focussed and on topic. Textbook title and date of publication along with unit numbers or pages being covered can also be part of the digest.

The teacher must determine what type of test will be developed in order to establish the amount of detail required in the Table. A main focus in teacher made assessments concerns students' cognitive abilities to understand and apply the concepts they have learned. There is less concern about the rapidity of a student's responses to questions than about the content of those responses. Accordingly, time limits on achievement tests are very generous, allowing all students enough time to consider each question and attempt to answer it. These tests are called power tests. Items on a power test have different levels of difficulty usually arranged in a hierarchy from knowledge level (easy) to increasing difficulty. A power test should be administered so that a very large percentage (90% is an acceptable minimum) of the pupils for whom it is designed will have

TABLE 1

Heading

Course Title: Art III
Grade level: 6, 7, 8,9, 10, **11**, 12 (Circle as appropriate)
Periods test is being used: 1 2 **3** 4 5 6 7 (Circle as appropriate)
Date of test: April 15, 2003
Subject matter digest: 19th and 20th Century Art. Includes artists from around the world. Oils and water colors as primary medium. Identify major works, styles, and schools.
Type Test: Power, Speed, **Partially Speeded** (Circle One)
Test Time: 45 minutes
Test Value: 100 points
Base Number of Test Questions: 75
Constraints: Test time, quantity of art available for test items

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	**Level**	**Instruct Time**	**Q/P/ %**								Q/P
1	Appl	95 16%	11/16	Matching		6(1)	5(2)				11/16
2	Comp	55 9%	7/10	MC		5(2)					5/10
3	Appl	50 8%	6/9	MC Essay	1(1)		2(2) 1(4)				4/9
4	Appl	35 6%	5/6	MC Essay	1(1)	1(1)	1(4)				3/6
5	Synth	45 8%	6/8	MC SA Essay			2(1)		2(1) 1(4)		5/8
6	Know	60 10%	8/10	True/ False MC	6(1) 1(2)						7/8
7	Appl	85 14%	10/14	MC	2(1)	2(1)	5(2)				9/14
8	Anal	60 10%	8/10	SA Essay				3(2) 1(4)			4/10
9	Comp	70 12%	9/12	Matching MC		6(1) 3(2)					9/12
10	Eval	40 7%	5/7	Essay						1 (7)	1/7
Total		600 min /100%	75/10 0		11/12	23/31	16/34	4/10	3/6	1/7	58/100

MC=Multiple Choice; SA = Short Answer **Q= Questions; P = Points**

ample time to attempt all of the items.

A speed test is one in which a student must, in a limited amount of time, answer a series of questions or perform a series of tasks of a uniformly low level of difficulty. The near-constant level of difficulty of the questions or tasks is such that, if the pupil had unlimited time, he or she could easily answer each question or perform each task successfully. The intent of a speed test is to measure the rapidity with which a pupil can do what is asked of him or her. Speed of performance frequently becomes important after students have mastered task basics as in using a keyboard, manipulatives, or phonics.

Tests are often a mixture of speed and power even when achievement level is the test's purpose. Such tests are called partially speeded tests. Teachers must check time limits carefully to be sure that all students will have the opportunity to address each test item adequately before the allotted time is up.

Once the purpose of the test as a power, speed or partially speeded test has been established, the teacher can decide the actual length of the test in minutes. The amount of time for the test is determined before test construction and is facilitated by using a Table of Specifications. Testing time, measured in minutes, is determined by a number of factors including: the number of objective to be tested; coverage of objectives; objective complexity; number of conditions to be tested; and levels of acceptable performance. In addition, teachers must look at students' age and ability levels, class time available, types of test items, length and complexity of test items, and amount of computation required.

Carey (1988) pointed out that the time available for testing depended not only on the length of the class period but also on students' attention spans. Completion of the test should be possible within one class period and the students should finish before they become fatigued (a six year old will not be able to take a 40 minute, paper-pencil test). A Table of Specifications insures that teachers will address all of these important issues in constructing an end of unit exam.

To continue our analogy, the something new at the wedding of teacher made tests and accountability is the use of an assessment plan to determine test value. The assessment plan has been around for a number of years but has not been associated with the development of a Table of Specifications. An assessment plan considers how many points the test is worth, how the test fits into the semester grade point total and eventually determines the Grade Point Average. An assessment plan determines total number of points available in a marking period. Semester and final grades for the year come from the six (or nine) week assessment plans added together.

The first step in developing an assessment plan is to list the assessment activities to be used in the class. The second step is to determine how many of each activity will be used in each grading period. The third step is to assign points according to the worth of the activity. This is a value judgment, e.g. "homework is less important than a unit exam but more important than answering questions in class." The following is an example of a six week assessment plan.

Example

Assessment Plan:
Determining Marking Period Point Values

Observation time on	
Objective/task	30 x 05 = 150
Homework	6 x 20 = 120
Class Participation	30 x 10 = 300
Quizzes	
Open book	3 x 10 = 30
Closed book	2 x 25 = 50
Tests	
Unit test	3 x 100 = 300
Marking period exam	1 x 200 = 200
Portfolio	0 for marking period
Total points marking period	1150

(Class work = 570 Tests = 580)

An assessment plan should be formed before each grading period begins. In the ex-

ample above, the points for testing and points for class work are evenly divided. This is the authors' point of view. Mehrens & Lehman (1973) suggested that the teacher determines the balance in the assessment plan. But, balance will not happen if there is inadequate planning. Adequate and extensive planning is required so that instructional objectives, the teaching strategy to be employed, the text material, and the evaluative procedures are all related in some meaningful fashion.

He also made suggestions for determining a base number of items to use per test. "Recall-level items require less time than application-level items, whatever the test format. Items that ask students to solve problems, analyze or synthesize information, or evaluate examples all require more time than do items that require students to remember a term, fact, definition, rule, or principle. Essay questions require more time than either selected-response or short-answer items" (p. 92).

Some rules of thumb exist for how long it takes most students to answer various types of questions:

- A true-false test item takes 15 seconds to answer unless the student is asked to provide the correct answer for false questions. Then the time increases to 30-45 seconds.
- A seven item matching exercise takes 60-90 seconds.
- A four response multiple choice test item that asks for an answer regarding a term, fact, definition, rule or principle (knowledge level item) takes 30 seconds. The same type of test item that is at the application level may take 60 seconds.
- Any test item format that requires solving a problem, analyzing, synthesizing information or evaluating examples adds 30-60 seconds to a question.
- Short-answer test items take 30-45 seconds.
- An essay test takes 60 seconds for each point to be compared and contrasted.

Fallback positions for determining how many questions should be on a test are how much time is available for testing and the level of performance required (test by conditions as well as action verb). In general, the more items on a test, the more valid and reliable the test will be. However, a test could be prohibitively long. On the other hand, a test with only one item per objective even if all items were answered correctly would provide insufficient evidence of proficiency. When all else fails look in the mirror to see who determines the number of test questions on a teacher made test.

Constraints are those variables that prevent testing in the manner that would be most appropriate for the level of instruction required to master the performance level indicated by the objective's action verb. Write the reason why you see a constraint, if there are no constraints state NONE. Types of constraints are time, personnel, cost, equipment, facilities, realism, logistics, communications, others.

The first heading in the body of the Table (see page 3) is called Learning Objectives. This heading has four subheadings: No; Level; Time; and Q/P/%. These subheadings, although distinct, are interrelated. No. represents the number designation of the objective. Either write the objective out in this space or put the number of the objective from an objective list in the space. If a list is used, it must be attached to the table.

The table itself is predicated on the writing of good performance objectives. A performance objective states the performance required or capability that is involved (action verb). The content is then specified through the behavior, situation, and special conditions components of the objective (condition{s}). When developing a Table you want to test all the objectives. You can only be sure students can perform the objectives which are tested. However, a constraint in doing that may be time. In that case you would want to do sampling of objectives.

You should sample among objectives only if it will solve a constraint problem. Document the sampling plan. Always test the most critical objectives. Test the less critical objectives in rotation randomly. Students are not informed of the objectives to be tested.

Sample among conditions if the action must be performed under each of two conditions develop items for each condition. If the action may be performed under either of two conditions, test the more difficult condition if only one can be tested. If the action must be performed under three conditions, test the two most critical ones. If the action must be performed under a large number of conditions, test at least 30% of them including the most critical ones.

Level equals domain level of the action verb of the objective. Level is assigning the objective's action verb to a category in Bloom's taxonomy. For example, Objective 1 is application and Objective 2 is comprehension. There are a number of lists of action verbs according to taxonomy level (e.g. Linn & Gronlund (2000), Appendix G). This assignment is done graphically so that you can look to the right of the assignment to see if there are any questions in levels beyond the assigned level. You can only test to the level taught. Otherwise you will be setting your students up for failure. You also must test objectives at full performance if you are going to state that students are competent at action verb level. At the level necessary, you can and should test the enabling skills for assurance that the students have the prerequisite skills to achieve full performance. In the following example from Table 1, partially reproduced here as Chart 1.

Objective 1 reads as follows "Identify architectural style in examples of 20th century revival style buildings around the world." There are no questions listed in the Table above application so we are not testing above the level taught. Under application there are five questions, therefore, Objective 1 is being tested at full performance. Under comprehension for Objective 1, there are six questions listed. These six questions test enabling skills required to obtain full performance. These questions may be such that examples of original styles of building architecture are presented and the student names them.

Bloom's Taxonomy's cognitive domain can be arranged in columns. Bloom's taxonomy is used because it provides the ability to develop a Table for a teacher made test in the cognitive, affective and psychomotor domains. The Tables used in this fastback as illustrations are all cognitive, however, the only difference between the cognitive and the affective and psychomotor is the interchange of the placement of the levels.

A Table ensures your test will include a variety of items at different levels of cognitive complexity. The cognitive domain is looked at as a set of steps. You must take the first step before you can attain the second, and so on. This mind set is very important when you look at congruency.

The example under LEVEL in Chart 1 illustrated an aspect of testing called CONGRUENCY. Congruency is teaching and testing at the same level. The level of the objective is matched with the placement of test items. Chart 2 is an example of congruency; testing what you are teaching using Objective 7 in Table 1.

Chart 1

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/ %								Q/P
1	Appl	95 16%	11/16	Matching		6(1)	5(2)				11/16

Chart 2

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/ %								Q/P
7	Appl	85 14%	10/14	MC	2(1)	2(1)	5(2)				9/14

Teaching	Application
Learning	Application
Test 1	Knowledge, Comprehension
Test 2	Comprehension, Application, Synthesis
Test 3	Version a. Knowledge, Application
	Version b. Knowledge, Comprehension, Application
	Version c. Comprehension, Application
	Version d. Application

The teacher is teaching Objective 7 at the application level. Similarly, to state that a student can fully perform at the application level, the test must assess at the application level. In the chart, if the teacher uses Test 1 Objective 7 has not been tested to the level of the objective, and you will not be able to state that the student who passed has mastered the objective. Test 2 is the reverse, you have set the students up for failure because you are testing at a mastery level you did not teach them to attain. Test 3 gives you a variety of ways to test for mastery of the objective level application, with Test 3 version b being used for Objective 7.

You would use Test 3 versions a, b, or c, if you were testing prerequisite or enabling objectives. While testing for maximum performance of the objective action verb you may need to ask questions on the prerequisite and enabling objectives to insure that the student had these abilities, otherwise you will not know why the student failed at the full performance measure. The testing of prerequisite and enabling objectives is extremely important, it helps you in being diagnostic and prescriptive in your test critique and determining if you taught with sufficient emphasis, depth, and breadth, the objective. An example of an enabling test question would be to give the value of $\prod$ if the objective full performance was to calculate the circumference of a circle given its radius.

To do the calculations for the TIME and Q/P/% columns of the table of specifications the teacher must use the following formulas for each objective in the table.

FORMULA "A"

time in class spent on objective (min) / total time for the instruction being examined (min) = % of instruction time

Example from Table 1 using Objective 1: total time for instruction 600 minutes. Time in class spent on Objective 1 95 minutes.

$$\frac{95}{600} = .16 \text{ or } 16\%$$

THEN the instructor should look at the number of test items and their point weight per question and complete Formula "B."

FORMULA "B"

point total of questions for objective / total points* on examination = % of examination value

Example from Table 1using Objective 1:

$$\frac{16}{100} = 16\%$$

Then the two percentages from Formula "A" and Formula "B" should be placed in Formula "C." If the outcome of Formula "C" is within the established parameters, the teacher may go to the next objective until they have completed the process for all objectives.

(*Total points is academic point value assigned to examination)

THEN the two percentages from Formula "A" and Formula "B" should be placed in Formula "C." If the outcome of Formula "C" is within the established parameters, the instructors may go to the next objective until they have completed the process for all objectives.

FORMULA "C"

Percent of instruction time = percent of examination value (within +- 2 percent, if not, redo test)

Example from Table 1using Objective 1:

16 = 16

Using as an example Table 1 objective NO. 1 had 95 minutes of instructional time spent on it. The total time of instruction covered by the test was 600 minutes. Using Formula "A" objective NO. 1 would have 16% of the instructional time. Using Formula "B" 16% of the instructional time would equate to 11 questions and 16 points. Formula "C" compares the two percentages. The percentages should be within the values established for content validity for an examination.

TIME equals the time, expressed in minutes, spent in class and other learning activities on the objective. Mehrens & Lehman (1973) state the major advantage of teacher made tests is that a teacher made test can be tailor made to fit the teacher's unique and/or particular objectives. However, the teacher must insure that appropriate weight is given during the test to those particular objectives. The formulas for calculating time have already been presented. Remember that all these times are in minutes and then converted to percent. The use of these formulas and their answers determine the distribution of numbers of questions on the test and point values assigned to said questions. Emphasis given during instruction must be used to assign weight in a test. Emphasis on an objective in a class and corresponding activities is a students' first and major clue to relevance and value of what is being taught. You have been in class where the teacher spend "X" amount f time on a subject and there is one question on the test covering that material and 14 on something that was covered by a paragraph in the text. The way the Table is constructed, time on objective, both direct and integrated is used to establish relevance of material to the students and for test construction. Total Time Spent Teaching all material is the baseline that is used to determine the weight given to the objective in the overall scheme of the Table. Mehrens & Lehman (1973) states there is no guarantee a "match" between instructional objectives and test item will take place if a Table is used; it will only indicate the number or proportion of test items to be allotted to each of the instructional objectives specified.

The final distribution of items in the Table of Specifications should reflect the emphasis given during the instruction. This concept of relative weight impacts both the construction of the Table and student perception that the test is fair. Objectives considered more important by the teacher should be allotted more test items. Similarly, areas of content receiving more instruction time should be allotted more test items. Too often students say, "I studied the chart in the book that we spent two days on and then there was nothing on the test.

And where did that essay on cause and effect come from." Relative weighting will alleviate these types of student comments.

Although the decisions involved in making the Table are somewhat arbitrary and the process is time consuming, the preparation of the Table of Specifications is one of the best means for ensuring that the test will measure a representative sample of instructionally related tasks.

The percentages are then used to determine the number of questions per objective and the value of points per objective.

Q/P/% is the number of questions (Q) and points (P) by percent (%) that represent the emphasis of instructional time based on relative weight. These are the number of questions and points that are the bench mark for test development. In the example below from Table 1, partially reproduced here as Chart 3, the Q/P/% of Objective 3 is in bold (6/9).

Linn & Gronlund (2000) provided the rationale behind the Q/P/% when they stated "We should like any assessment of achievement that we construct to produce results that represent both the content areas and the objectives we wish to measure, and the table of specifications aids in obtaining a sample of tasks that represents both. The percentages in the table indicate the relative degree of emphasis that each content area and each instructional objective is to be given in the test" (p. 80).

Linn & Gronlund (2000) further stated "this table indicates both the total number of test items and assessment tasks and the percentage allotted to each objective and each area of content. For classroom testing, using the number of items may be sufficient, but the percentages are useful in determining the amount of emphasis to give to each area" (p. 562).

Chart 3

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/%								Q/P
3	Appl	50 8%	**6/9**	MC Essay	1(1)		2(2) 1(4)				4/9

Chart 4 Test Item Format

Type of Test Item	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Multiple Choice	X	X	X	X	X	
Matching	X	X	X	X		
True-False	X	X				
Short Answer		X	X	X	X	
Essay		X	X	X	X	X

Chart 5

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/%								Q/P
1	Appl	95 16%	11/16	Matching		6(1)	5(2)				11/16
2	Comp	55 9%	7/10	MC		5(2)					5/10
3	Appl	50 8%	6/9	MC Essay	1(1)		2(2) 1(4)				4/9
5	Synth	45 8%	6/8	MC SA Essay			2(1)		2(1) 1(4)		5/8

Linn & Gronlund (2000) summed up Q/P/% when they stated "the final distribution of items in the table of specifications should reflect the emphasis given during the instruction. Objectives considered more important by the teacher should be allotted more test items. This applies not only to the items on the classroom test but also to performance assessment tasks. The weight given to the performance of such assessment tasks should reflect the importance of the objective. Similarly, areas of content receiving more instruction time should be allocated more test items and assessment tasks" (p. 147).

The second major heading in the Table body is ITEM TYPE. Item type is the type(s) of test item(s) used to test the student's ability to obtain the objective. The Test Item Format Chart below provides a visual representation of the levels of the cognitive domain that can be tested by the five basic test items used on teacher made tests. Depending on complexity, wherever possible use the most simplistic test item format.

Using Table 1, partially reproduced here as Chart 5, as an example, Objectives 1 and 3 are both full performance at the application level. However, they are being tested by different item types, but with the correct types of questions as prescribed by the chart. The use of the essay in Objective 3 may be to explain reasoning or a procedure required by the objective for full performance.

The third subheading in the Table body is Bloom's Taxonomy/Congruency. LEVELS of the domain tested and the total number of the types of questions in the level(s) tested are listed. This will assist in determining if testing is at multiple levels, only at the highest level, or at too high a level. The base Table of Specifications is set up for the cognitive domain. If testing the affective or psychomotor domain, the Table is the same, except the cognitive levels would be replaced by the levels of the affective or psychomotor domain.

TOTAL in a row equals the number of questions testing an objective (Table 1, partially reproduced here as Chart 6, Objective 6). Total in a column equals the number of questions testing a domain level (Chart 6, Application level: 16/34).

Chart 6

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/ %								Q/P
1	Appl	95 16%	11/16	Matching		6(1)	**5(2)**				11/16
2	Comp	55 9%	7/10	MC		5(2)					5/10
3	Appl	50 8%	6/9	MC Essay	1(1)		**2(2) 1(4)**				4/9
4	Appl	35 6%	5/6	MC Essay	1(1)	1(1)	**1(4)**				3/6
5	Synth	45 8%	6/8	MC SA Essay			**2(1)**		2(1) 1(4)		5/8
6	Know	60 10%	8/10	True/ False MC	**6(1)** **1(2)**						**7/8**
7	Appl	85 14%	10/14	MC	2(1)	2(1)	**5(2)**				9/14
8	Anal	60 10%	8/10	SA Essay				3(2) 1(4)			4/10
9	Comp	70 12%	9/12	Matching MC		6(1) 3(2)					9/12
10	Eval	40 7%	5/7	Essay						1 (7)	1/7
Total		600 min /100%	75/10 0		11/12	23/31	**16/ 34**	4/10	3/6	1/7	**58/ 100**

The sums of the columns and row should be equal. If they are not, then the addition is incorrect. The bottom right hand corner is where the column and row totals are found (Chart 6: 56/100). The total number of questions for each level of the domain is summed objective (Chart 6, Objective 6: 7/8). Then all the levels of the domains are added. This total should equal the total number of questions which where determined to be on the test. Similarly the values of each question for each objective are summed and the total of all points is added. This total should equal the set value of the examination (Chart 6: 56/100; testing heading stated test value was 100 points, they match).

NOTE: Common sense is important. Make point values whole numbers, no 1.5, etc. You will spend too much time grading. The questions per objective and point value are assigned based on percent of time taught including direct instruction and integrated instruction. Therefore one percent equals one question worth one point. However, if you use a question and it is worth two points look at that as two questions. If you have an essay question worth 5 points look at it as five questions. Also, when rounding up or down to get a full question or point, always round up for the higher level objectives. Number of questions per objective can go down but point value per objective is not changed.

Using Table 1, partially reproduced here as Chart 7, the objectives and points are:

Summarizing the objectives and their point totals in Chart 7 would look like this:

Objectives	#	Point Value	
10%	1	Knowledge	12
21%	2	Comprehension	21
44%	4	Application	34
10%	1	Analysis	10
08%	1	Synthesis	06
07%	1	Evaluation	07
100%			100

To check that your test is assessing as taught you look at the total row at the bottom of the Table1, partially reproduced here as Chart 8 and you will see if values are within line.

To keep with the wedding theme something needs to be borrowed. We have borrowed two things for this wedding. We are going to borrow from Carey (1988) some thoughts on how to make the Table provide a test that is both valid and reliable.

Carey (1988) stated "During the design of classroom tests, you need to be concerned with the validity and reliability of test scores. We have discussed content validity and how the Table will provide for it. Reliability is not normally associated with the Table. Reliability refers to the consistency or stability of scores obtained from a test. If the scores

Chart 7

Learning Objective			
No	**Level**	**Instruct Time**	**Q/P/ %**
1	Appl	95 16%	11/16
2	Comp	55 9%	7/10
3	Appl	50 8%	6/9
4	Appl	35 6%	5/6
5	Synth	45 8%	6/8
6	Know	60 10%	8/10
7	Appl	85 14%	10/14
8	Anal	60 10%	8/10
9	Comp	70 12%	9/12
10	Eval	40 7%	5/7

Chart 8

Learning Objective				Item Type	Bloom's Taxonomy/Congruency						
					Know	Comp	Appl	Anal	Syn	Evl	Total
No	Level	Instruct Time	Q/P/ %								Q/P
Total		600 min /100%	75/ 100		11/12	23/31	16/34	4/10	3/6	1/7	58/100

are unreliable, decisions or inferences based on them are dubious. Tests must be designed carefully to yield reliable and valid scores" (p. 95).

Carey (1988) continued that there are "five steps during the design stage you must take to achieve reliable test results: (1) select a representative sample of objectives from the goal framework; (2) select enough items to represent adequately the skills required in the objective; (3) select item formats that reduce the likelihood of guessing; (4) prescribe only the number of items students can complete in the time available; and (5) determine ways to maintain positive student attitudes toward testing. The subordinate skills in an instructional goal framework should be divided into prerequisite skills (skills students should have mastered before entering a unit of instruction) and enabling skills (skills that comprise the main focus of instruction for a unit)" (p. 95). The Table presented takes into account the five steps that will make a test reliable.

The second thing borrowed is Linn & Gronlund's (2000) idea to embed related non-test assessment procedures in an expanded Table of Specifications.

Reproducing the assessment plan shown earlier (page 5) as Chart 8 and we could add the class attendance, homework, class participation, and quiz points used during the instructional time that our test covered. In Table 1 (reproduced from page 3) with the added non-test points we have added the categories and values in the heading of the Table and then emphasized in the body of the Table the non-test learning activities and their relative points by underlining.

EXAMPLE

Assessment Plan: Determining Marking Period Point Values

Observation time on objective/task	30 x 05	= 150
Homework	6 x 20	= 120
Class Participation	30 x 10	= 300
Quizzes		
Open book	3 x 10	= 30
Closed book	2 x 25	= 50
Tests		
Unit test	3 x 100	= 300
Marking period test	1 x 200	= 200
Portfolio	0 for marking period	
Total points marking period	1150 (Class work = 570 Tests = 580)	

Figure 1

	Objectives						
	Knows		Understands	Interprets	Skill in		
Content	*Symbols and Terms*	*Specific Facts*	*Influence of Each Factor on Weather Formation*	*Weather Maps*	*Using Measuring Devices*	*Constructing Weather Maps*	*Total Number of Items*
Air pressure	2	3	3	3	Observe pupils using measuring devices (rating scale)	Evaluate maps constructed by pupils (checklist)	11
Wind	1	2	8	2			16
Temperature	2	2	2	2			8
Humidity and precipitation	2	1	2	5			10
Clouds	2	2	1				5
Total number of items	12	10	16	12			50
Percentage of evaluation	12%	10%	16%	12%	25%	25%	100%

Summary

The Table of Specifications is used to show two things; first, the emphasis of the test item is equal to the emphasis of the instructional time. Instructors are testing what they taught. The second thing a Table shows is the test is assessing at the appropriate level(s). If there are constraints, always test at the highest level. If an individual can perform the most difficult aspects of the objective, the instructor can "assume" the lower levels can be done. However, if testing the lower levels, the instructor cannot "assume" the individual can perform the higher levels. If there are no constraints, testing across the levels can be conducted so as to indicate where a student or class erred when they did not perform at the highest level.

Teacher made tests and accountability have been united. It takes effort to make a marriage work just as it does to make a teacher made test meet the validity and reliability requirements of accountability. The Table of Specifications is the tie that binds.

References

Bennett, D. T., Wesley, H., & Dana-Wesley, M. (1999). Planning for imminent change in college admissions: Research on alternative admissions criteria. *Journal of College Student Retention, 1*(1), 83-92.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning.* New York: McGraw-Hill, Inc.

Carey, L. M. (1988). *Measuring and Evaluating School Learning.* Boston: Allyn, and Bacon, Inc.

Cool, L. C. (2002). Testing testing...testing... *Good Housekeeping, 235*(2), 83-87.

Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*(2), 179-194.

Dills, C. R. (1998). The table of specifications: A tool for instructional design and Development. *Educational Technology, 38*(3), 44-51.

Ediger, M. (2002). *Assessing state mandated tests.* (Report No.TM033724) (ERIC Reproduction Service No. ED463297)

Eisenberg, M., & Serim, F. (2002). No child left behind - the implications for Educators. *Multimedia Schools, 9*(4), 27-29.

Falk, B. (2002). Standards-based reform: Problems and Possibilities. *Phi Delta Kappan, 83*(8), 612-620.

Gentry, D. L. (1989). *Teacher-made test construction.* (Report No. TM014285). Paper presented at the Annual Meeting of the Mid-south Educational Research Association (Littlerock, Arkansas, November 8-10, 1989). (ERIC Reproduction Service No. ED313444)

Gredler, M. E. (1999). *Classroom Assessment and Learning.* New York: Longman, an imprint of Addison Wesley Longman, Inc.

Halpern, D. F. (2002). Sex differences in achievement scores: Can we design assessments that are fair, meaningful, and valid for girls and boys? *Issues in Education, 8*(1), 2-21.

Hoover, E. (2002a). SAT is set for an overhaul, but questions linger about the test. *Chronicle of Higher Education, 48*(38), A35-A36.

Hoover, E. (2002b). College Board approves major changes for the SAT. *Chronicle of Higher Education, 48*(43), A34-A35.

Imber, M. (2002). The problem with grading. *American Schools Board Journal, 189*(6), 40-41, 47.

Jenkins, N. J. (1992). *The Scholastic Aptitude Test as a predictor of academic success: A literature review.* (Report No. TM019236) (ERIC Reproduction Service No. ED354243)

Kubiszyn, T., & Borich, G. (2003) *Educational Testing and Measurement: Classroom Application and Practice.* (7th ed.) New York: John Wiley & Sons, Inc.

Lei, P. W., Bassiri, D., & Schultz, E. M. (2001). *Alternatives to the grade point average as a measure of academic achievement in college.* (Report No. TM033669) American College Testing Program. (ERIC Reproduction Service No. ED462407)

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching.* (8th ed.) Columbus, OH: Merrill. And imprint of Prentice Hall.

Marshall, J. (1997). Seniors boost higher grades, so-so SAT scores. *Christian Science Monitor, 89*(198), 12.

Martin, R. (1988). *Developing and Improving the Quality of Written Tests.* (Report No.

CE050757) Paper presented at the Midwest Nuclear Training Association's Annual Nuclear, Columbus, Ohio, September 22, 1988. (ERIC Reproduction Service No. ED298264)

Mehrens, W. A., & Lehman, I. J. (1973). *Measurement and Evaluation in Education and Psychology.* (4th ed.) Chicago: Holt, Rinehart and Winston, Inc.

Micceri, T. (2001). *Facts and Fantasies regarding admission standards.* Report No. HE034083. Paper presented at the Annual Meeting of the Association for Institutional Research (Long Beach, CA, June 3-6, 2001. (ERIC Reproduction Service No. ED453757)

Murphy, M. (1981). What's your classroom testing validity quotient? *School Shop, 40*(5), 18-20, 27.

Nathan, J. (2000). Taking on the NCAA. *Phi Delta Kappan, 82*(4), 310-312.

Newell, R. J. (2002). A Different look at accountability: The edvisions approach, *Phi Delta Kappan, 84*(3), 208-211.

Notar, C. E., & Hoffman, S. (1994) U*nited States Army Military Police School Test Regulation,* Fort McClellan, Alabama.

Oosterhof, A. (2002). *Classroom Applications of Educational Measurement.* (3rd ed.) Columbus, OH: Merrill Prentice Hall.

Patton, T. K. (1998). *Differential prediction of college performance between gender*, (Report No. TM029407) Middle Tennessee State University.(ERIC Reproduction Service No. ED427030)

Perfetto, G. (2002). Predicting academic success in the admissions process: Placing an empirical approach in a larger process. *College Board Review*, (196), 30-35.

Stumpf, H., & Stanley, J. C. (2002). Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement, 62*(6), 1042-1052.

Weis, S. F. (1976). The use of the table of specifications in developing educational objectives and evaluation. Illinois *Teacher of Home Economics, 19*(3), 167-169.