

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to the *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 18, Number 3, February 2013

ISSN 1531-7714

Classroom Test Construction: The Power of a Table of Specifications

Helenrose Fives & Nicole DiDonato-Barnes
Montclair State University

“But we only talked about Grover Cleveland for – like 2 seconds last week. Why would she put that on the exam?”

“You know how teachers are... they’re always trying to trick you.”

“Yeah, they find the most nit-picky little details to put on their tests and don’t even care if the information is important.”

“It’s just not fair. I studied everything we discussed in class about the Gilded Age and the things she made a big deal about, like comparing the industrialized north to the agriculture in the south. I really thought I understood what was going on – how the U.S. economy and way of life changed with industry, railroads, and unions. And to think all she asked was ‘What was the South’s economic base!’ Oh and ‘What were Grover Cleveland’s terms as president?’ Really? Grrr.”

As a student have you ever felt that the test you studied for was completely or partially unrelated to the class activities you experienced? As a teacher have you ever heard these complaints from students? This is not an uncommon experience in most classrooms. Frequently there is both a real and perceived mismatch between the content examined in class and the material assessed on an end of chapter/unit test. This lack of coherence leads to a test that fails to provide evidence from which teachers can make valid judgments about students’ progress (Brookhart, 1999). One strategy teachers can use to mitigate this problem is to develop a Table of Specifications (TOS).

What is a Table of Specifications?

A TOS, sometimes called a test blueprint, is a table that helps teachers align objectives, instruction, and assessment (e.g., Notar, Zuelke, Wilson, & Yunker, 2004). This strategy can be used for a variety of assessment methods but is most commonly associated with constructing traditional summative tests. When constructing a test, teachers need to be concerned that the test measures an adequate sampling of the class content at the cognitive level that the material was taught. The TOS can help teachers map the amount of class time spent on each objective with the cognitive level at which each objective was taught thereby helping teachers to identify the types of items they need to include on their tests. There are many approaches to developing and using a TOS advocated by measurement experts (e.g., Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, & Wittrock, 2001; Gronlund, 2006; Reynolds, Livingston, & Wilson, 2006).

In this article, we describe one approach to using a TOS developed for practical classroom application. Our approach to the TOS is intended to help classroom teachers develop summative assessments that are well aligned to the subject matter studied and the cognitive processes used during instruction. However, for this strategy to be helpful in your teaching practice, you need to make it your own and

consider how you can adapt the underlying strategy to your own instructional needs. There are different versions of these tables or blueprints (e.g., Linn & Gronlund, 2000; Mehrens & Lehman, 1973; Nortar et al., 2004), and the one presented here is one that we have found most useful in our own teaching. This tool can be simplified or complicated to best meet your needs in developing classroom tests.

What is the Purpose of a Table of Specifications?

In order to understand how to best modify a TOS to meet your needs, it is important to understand the goal of this strategy: improving validity of a teacher's evaluations based on a given assessment. **Validity** is the degree to which the evaluations or judgments we make as teachers about our students can be trusted based on the quality of evidence we gathered (Wolming & Wilkstrom, 2010). It is important to understand that validity is not a property of the test constructed, but of the inferences we make based on the information gathered from a test. When we consider whether or not the grades we assign to students are accurate we are questioning the validity of our judgment. When we ask these questions we can look to the kinds of evidence endorsed by researchers and theorists in educational measurement to support the claims we make about our students (AERA, APA, NCME, 1999). For classroom assessments two sources of validity evidence are essential: evidence based on test content and evidence based on response process (APA, AERA, NCME, 1999). At the beginning of this article the students complained about a lack of coherence between the subject matter discussed in class (test content evidence) as well as the kind of thinking required on the test (response process evidence).

Test content evidence was questioned by the first student who stated "*But we only talked about Grover Cleveland for—like 2 seconds last week...*" In this comment the student is concerned that the material (content) he studied and the teacher emphasized was not on the test. **Evidence based on test content** underscores the degree to which a test (or any assessment task) measures what it is designed (or supposed) to measure (Wolming & Wilkstrom, 2010). If an Algebra I teacher gave an exam on the proof of Pythagoras' theorem and

based her Algebra I grades on her students' response to that exam, most of us would argue that the exam and the grades were unjustified. In assessment we would say that her judgment lacked evidence of test content agreement, because the evidence used (data from a geometry test) to make the judgment did not reflect students' understanding of the targeted content (algebra). Your classroom tests must be aligned to the content (subject matter) taught in order for any of your judgments about student understanding and learning to be meaningful. Essentially, with test-content evidence we are interested in knowing if the measured (tested/assessed) objectives reflect what you claim to have measured.

Response process evidence is the second source of validity evidence that is essential to classroom teachers. **Response process evidence** is concerned with the alignment of the kinds of thinking required of students during instruction and during assessment (testing) activities. For example, the last student in the opening scenario implied that class time was spent comparing the U. S. North and South during the Gilded Age (circa 1877-1917) yet on the test the teacher asked a low level recall question about the economic base of the South. The inclusion of a question such as this is supported by evidence of test-content, the student recalled the topic mentioned. But the depth of processing required to compare the North and South during instruction involved more attention and deeper understanding of the material. This last student clearly felt that there was a lack of congruence in the kind of thinking required for this test and during instruction.

Sometimes the tests teachers administer have evidence for test content but not response process. That is, while the content is aligned with instruction the test does not address the content at the same depth or level of meaning that was experienced in class. When students feel that they are being tricked or that the test is overly specific (nit-picky) there is probably an issue related to response process at play. As test constructors we need to concern ourselves with evidence of response process. One way to do this is to consider whether the same kind of *thinking* is used during class activities and summative assessments. If the class activity focused on memorization then the final test

should also focus on memorization and not on a thinking activity that is more advanced.

Table 1 provides two possible test items to assess the understanding of sources of validity evidence. In Table 1, Option A assesses whether or not students can *recognize* a definition of test content validity evidence. Option B assesses whether or not students can *evaluate* the prompt and *apply* the type of validity evidence described in the scenario. Thus, these two items require different levels of thinking and understanding of the same content (i.e., recognizing vs. evaluating/applying). Evidence of response process ensures that classroom tests assess the level of thinking that was required for students during their instructional experiences.

Table 1: Examples of items assessing different cognitive levels

Option A

The degree to which the test assesses the appropriate content material it intends to measure refers to evidence of:

- a. test content.
- b. response process.
- c. criterion relationships.
- d. test consequences.

Option B

Constance is fed up with Mr. Kent, her history teacher. He asks the most obscure items on his test about things that were never discussed in class!

What kind of test evidence is Constance concerned about?

- a. Test Content
- b. Response Process
- c. Criterion Relationships
- d. Test Consequences

Levels of thinking. Six levels of thinking were identified by Bloom in the 1950's and these levels were revised by a group of researchers in 2001 (Anderson et al). Thinking that emphasizes recall, memorization, identification, and comprehension, is typically considered to be at a lower level. Higher levels of

thinking include processes that require learners to apply, analyze, evaluate, and synthesize.

Table 2 presents two released questions from a 5th grade U. S. History test on the Middle Colonies. Take a moment to review the two test items. The first item is written to assess student thinking at a lower level because it asks the student to recall facts and identify the same facts in the answer choices given. This question does not require students to do more than repeat the information presented in the textbook. In contrast, the second item addresses similar content but is written to assess higher levels of thinking. This item requires students recall information about Maryland colonists and apply that information to the examples given.

Table 2: Examples of a lower- and higher-level items

Item	Cognitive Level
1. Maryland was settled as a/an	<i>Lower level.</i> This item requires students to demonstrate recall
a. area to grow rice and cotton.	knowledge of Maryland settlers. This is a direct recall item that does not
b. safe place for English debtors.	require analysis or
c. colony for indentured servants.	application.
d. refuge for Roman Catholics.	
2. Which of the following people would <i>most</i> want to settle in Maryland?	<i>Higher Level.</i> This question requires students to apply what they know about the colony of Maryland,
a. A Catholic from southern England.	analyze each of the item options as potential
b. A debtor from an English Prison.	Maryland settlers.
c. A tobacco planter.	
d. A French trapper.	

When considering test items people frequently confuse the type of item (e.g., multiple choice, true false, essay, etc.) with the type of thinking that is needed to respond to it. All types of item formats can

be used to assess thinking at both high and low levels depending on the context of the question. For example an essay question might ask students to “Describe four causes of the Civil War.” On the surface this looks like a higher level question, and it could be. However, if students were taught “The four causes of the Civil War were...” verbatim from a text, then this item is really just a low-level recall task. Thus, the thinking level of each item needs to be considered in conjunction with the learning experience involved. In order for teachers to make valid judgments about their students’ thinking and understanding then the thinking level of items need to match the thinking level of instruction. The Table of Specifications provides a strategy for teachers to improve the validity of the judgments they make about their students from test responses by providing content and response process evidence.

Using a Table of Specification to Support Validity

The TOS provides a two-way chart to help teachers relate their instructional objectives, the cognitive level of instruction, and the amount of the test that should assess each objective (Nortar et al., 2004). Table 3, illustrates a modified TOS used to develop a summative test for a unit of study in a 5th grade Social Studies class. The TOS provides a framework for organizing information about the instructional activities experienced by the student. Take a few moments to review the TOS. Be aware that before the teacher can construct the TOS, he/she will need to determine (1) the number of test items to include and (2) the distribution of multiple choice and short answer items. In the following example, the teacher has decided to include 10 items (i.e., 7 multiple choice and 3 short answer). The TOS provided here is simplified by limiting the levels of cognitive processing to high and low levels, rather than separating out across the six levels of cognitive processing identified by Bloom (1956) and updated by Anderson et al (2001). We do this for practical reasons, it is difficult to parse out test items by each level and teachers have limited time to engage in these activities. Furthermore, using this broader classification ameliorates the philosophical criticisms about the hierarchical nature of the

taxonomy and the distinction among the categories (Kastberg, 2003).

Evidence for test content.

One approach to gathering evidence of test content for your classroom tests is to consider the amount of actual class time spent on each objective. Things that were discussed longer or in greater detail should appear in greater proportion on your test. This approach is particularly important for subject areas that teach a range of topics across a range of cognitive levels. In a given unit of study there should be a direct relation between the amount of class time spent on the objective and the portion of the final assessment testing that objective. If you only spent 10% of the instructional time on an objective, then the objective should only count for 10% of the assessment. A TOS provides a framework for making these decisions.

A review of Table 3 reveals a 7 column TOS (labeled A-G). The information in columns A, B, and C are taken directly from the teacher’s lesson plans and reflective notes. Using a TOS helps teachers to be accountable for the content they teach and the time they allocate to each objective (Nortar et al., 2004). The numbers in Column D are the result of a percentage calculation. These numbers reflect the percent of total class time for the unit of study that was spent on each objective. To determine the percentage of total class time that was spent on each objective you take the minutes spent on the objective (column C) divided by the total minutes (bottom of column C), multiplied by 100. For instance the last objective in the table was allocated 10% of the overall class time (15 minutes/150 minutes of total instruction * 100).

How many items should be on your test?

In the top of Column E of Table 3, you should note that for this test the teacher has decided to use 10 items. The number of items to include on any given test is a professional decision made by the teacher based on the number of objectives in the unit, his/her understanding of the students, the class time allocated for testing, and the importance of the assessment. Shorter assessments *can* be valid, provided that the assessment includes ample evidence on which the teacher can base inferences about students’ scores.

Table 3: A Sample Table of Specifications for Fifth Grade Social Studies Chapter 6: The Middle Colonies

A	B	C	D	E	F	G
	Instructional Objectives	Time Spent on Topic (minutes)	Percent of Class Time on Topic	Number of Test Items: 10	Lower Levels -Knowledge -Recall -Identification -Comprehension	Higher Levels -Application -Analysis -Evaluation -Synthesis
Day 1	1. Identify the various groups who settled the Middle Atlantic Colonies.	15	10.00%	1.00	1 Multiple Choice	
	2. Summarize the contributions of different religious and cultural groups to the settlement of the Middle Atlantic Colonies.	15	10.00%	1.00	1 Short Answer	
Day 2	3. Identify George Whitefield as an early leader of the Great Awakening	10	6.70%	.67	1 Multiple Choice	
	4. Evaluate the impact of the Great Awakening sermons on English colonists.	20	13.30%	1.33		1 Multiple Choice
Day 3	5. Describe the physical features that helped Philadelphia become a main port.	15	10.0%	1.00	1 Multiple Choice	
	6. List ways in which immigrants aided Philadelphia's growth and prosperity.	10	6.70%	.67	1 Short Answer	
	7. Identify the contributions Benjamin Franklin made to Philadelphia.	5	3.30%	.33	---	
Day 4	8. Interpret information in a circle graph.	15	10.00%	1.00		1 Multiple Choice
	9. Gather and organize information using a circle graph.	15	10.00%	1.00		1 Short Answer
Day 5	10. Identify the challenges faced by backcountry settlers.	5	3.30%	.33	---	
	11. Analyze the importance of the Great Wagon Road as an early transportation route.	10	6.70%	.67		1 Multiple Choice
	12. Explain how backcountry settlers adapted to and made use of the resources available to them.	15	10.00%	1.00		1 Multiple Choice
		150	100.00%	10	5	5

Typically, because longer tests can include a more representative sample of the instructional objectives and student performance, they generally allow for more valid inferences. However, this is only true when test items are good quality. Furthermore, students are more likely to get fatigued with longer tests and perform less

well as they move through the test. Therefore, we believe that the ideal test is one that students can complete in the time allotted, with enough time to brainstorm any writing portions, and to check their answers before turning in their completed assessment. The creator of the TOS in Table 3 decided to create a

10 item test that would include 7 multiple choice items and 3 short answer items. Just a reminder, this is a professional decision made by the teacher based on his/her knowledge of the students, the classroom context, and the role of this test in relation to other assessments for the grade period.

The remainder of column E is used to determine how many test items (of equal value) should be used to assess each objective. To make this calculation you simply multiply the percentage of the test each objective should assess by the number of items the teacher has decided to include on the test. So for the first objective you multiply $10\% \times 10 \text{ items} = 1 \text{ item}$. An alternative approach to this step of the TOS is to think about the number of points the test is worth. If the test is worth 10 points of a student's total grade then one point of this test should assess objective 1. The use of points allows for varied weights to be applied to items that assess different objectives. However, in practice this can sometimes create more confusion than it does a quality assessment.

By now you may have noticed that the number of items per objective (Column E) does not always come out to a nice even number. In these cases the teacher must again use his/her professional judgment to decide whether to assess objectives with partial point values or not. In this example the teacher chose to "round up" the items for objectives 3, 6, and 11 and "round down" the items for objectives 4, 7, and 10. This brings up an important point about constructing classroom tests. Every objective does not need to be assessed in every assessment. A TOS can help you make sure that the most relevant objectives are assessed and that a sampling of less prominent ones are also included. A student when preparing for a test studies everything and gains an understanding of the content. What can actually be assessed is only a sampling of the students' knowledge at a particular point.

Evidence for response process.

Columns F and G indicate the professional judgment of the teacher. Based on the number of items per objective as calculated in Column E, the teacher must now decide which objectives to assess and with how many items. The teacher must also decide whether the objective should be tested at a low or high level

based on the learning objective and how the content was taught. If you look at the first objective "Identify the various groups who settled the Middle Atlantic Colonies" the teacher determined that this should be included on the test – 10% of the total class time was spent on this objective and thus 10% (or one item) of the test should assess this objective. The teacher has indicated that he/she will select or construct one multiple choice item to assess this at a lower cognitive level that will require the student to identify or recall or recognize the correct answer.

As mentioned above the teacher must decide which type of question to use to assess each objective at the correct level. When making this decision a teacher should consider the best way to get the desired information from the student. For instance, in Table 3, the teacher has indicated that he/she will use a short answer question to assess objective 2 "Summarize the contributions of different religious and cultural groups to the settlement of the Middle Atlantic Colonies." While this is considered a low level thinking item, simply rephrasing the material taught in class, it lends itself to a short answer item because students are required to put these descriptions in their own words rather than just selecting from a series of choices. This may prove to be a more challenging item for fifth grade students because it requires them to recall and write out their responses. However, the thinking involved in this task is still low level. In contrast, if the students were asked to make comparisons or evaluations between the groups then the objective would be at a higher level.

The TOS is a Tool for Every Teacher

The cornerstone of classroom assessment practices is the validity of the judgments about students' learning and knowledge (Wolming & Wilkstrom, 2010). A TOS is one tool that teachers can use to support their professional judgment when creating or selecting test for use with their students. The TOS can be used in conjunction with lesson and unit planning to help teacher make clear the connections between planning, instruction, and assessment.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition)*. New York: Longman.
- Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practices*, 18, 5-13.
- Gronlund, N. E. (2006). *Assessment of student achievement*. (8th ed.). Boston: Pearson.
- Kastberg, S. E. (2003). Using Bloom's taxonomy as a framework for classroom assessment. *The Mathematics Teacher*, 96, 402.
- Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Columbus, OH: Merrill.
- Mehrens, W. A. & Lehman, I. J. (1973). *Measurement and evaluation in education and psychology*. Chicago: Holt, Rinehart, and Winston, Inc.
- Notar, C.E., Zuelke, D. C., Wilson, J. D. & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, 31, 115-129.
- Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). *Measurement and Assessment in Education*. Pearson: Boston.
- Wolming, S. & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, 17, 117-132.

Citation:

Fives, Helenrose & DiDonato-Barnes, Nicole (2013). Classroom Test Construction: The Power of a Table of Specifications. *Practical Assessment, Research & Evaluation*, 18(3). Available online: <http://pareonline.net/getvn.asp?v=18&n=3>

Corresponding Author:

Helenrose Fives
Department of Educational Foundations
Montclair State University
Montclair, New Jersey 07043.
Email: fivesh [at] mail.montclair.edu