MEXTESOL Journal, Vol. 16, No. 1, Spring, 1992

Exam Development: The First Step

JoAnn Miller

Instituto Mexicano Norteamericano de Relaciones Culturales, A.C.
Mexico City

Probably one of a teacher's least favorite tasks is the development of written tests. Whether you believe written tests are necessary or not, they are a part of our lives. Students expect tests, parents demand tests and the administration often judges the success of language programs based on the results of tests. In reality, a written test can be very useful for both the student and the classroom teacher: For the student, by fomenting feedback and for the classroom teacher, by allowing a close examination of the techniques which worked and of those which should be adapted in the future and of how well the students have assimilated what we have been teaching.

Probably one of the most difficult steps in test writing is deciding exactly what to test. Unfortunately many tests and exams do not have *content validity*. An examination is said to have *content validity* "if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned." (Hughes 1989: 22) This means if a test is not carefully planned and developed it might be testing something that was not really taught in the classrooms; this is a very unfair situation for students who usually assume they will be tested on what they have studied.

Careful planning before writing an exam can improve content validity. Arthur Hughes states that "the essential first step in testing is to make oneself perfectly clear about what it is one

wants to know and for what purpose." He suggests the test writer consider the answers to the following questions before beginning to write: (*Note: Superscripts refer to the notes at the end of the article and are not part of the original quotation.*)

> *--What kind of test is it to be? Achievement*[1] *(final or progress), proficiency*[2]*, diagnostic*[3]*, or placement*[4]*?*
>
> *--What is its precise purpose?*
>
> *--What abilities are to be tested?*
>
> *--How detailed must the results be?*
>
> *--How accurate must the results be?*
>
> *--How important is backwash?* [5]
>
> *--What constraints are set by unavailability of expertise, facilities, time (for construction, administration and scoring)?*
>
> *(Hughes 1989: 48)*

---

[1] *Achievement tests* are used to measure the extent of learning in a specific course. They could be monthly tests, unit exams, midterm, semester or final examinations.

[2] *Proficiency Exams* are global measures of ability. They are not usually related to a specific course and are often used to select candidates for specific jobs or study programs.

[3] *Diagnostic Tests* are often used by classroom teachers to find out exactly which problems a group of students might have before beginning a course. They are used to plan future reviews and course content.

[4] *Placement Tests* are used to put new students in particular courses. They are similar to proficiency exams, but they are not as general since they are designed with a specific program in mind.

[5] *Backwash* is the effect of testing on the teacher and the learner. It can be positive (The test can be a valuable learning experience in itself.) or negative (It might not directly relate to the goals of the learning experience or it might be seen as useless or unfair.)

In general, classroom teachers write only achievement exams; special committees are usually formed or commercially available exams are used to fulfill the other needs. However, no matter what purpose the exam to be developed will have, one aspect is very important: the exam writer or writers must have a clear idea of the precise purpose and make-up of the exam. If exam writers are unclear as to an exam's purpose or content or if a team of writers is not in agreement, the resulting exam will reflect this confusion. Usually the purpose of an achievement test is to measure how much a given student has learned in a course. But, what is an acceptable level of mastery for a particular course? Should the student understand and be able to use everything that was seen in the course or is seventy or eighty percent mastery sufficient? Is the exam going to test only grammatical ability or is reading comprehension to be included? Writers must also agree on how detailed and how accurate the results must be. How important is the exam? A weekly quiz might not need as much time devoted to its preparation as a semester or final exam. Backwash should also be considered. Will the exam be a learning experience or will the students see it as a useless task taken only to get a grade? Exam writers must agree on basic philosophical questions before they begin writing.

Also the writers must be realistic. Not all teachers giving the exam are equally prepared and the exam-taking circumstances are not always ideal. Can all the teachers read that wonderful listening comprehension passage clearly? And, even if they can, is the room quiet enough for the students to hear it adequately?

Timing is also important. Allow enough time to develop an exam and, ideally, pretest it on an isolated group of students

before administering it formally. An exam should be ready more than a week or two before giving it. Time is necessary to plan the exam, to proofread it, and to print, collate and distribute the copies.

For one teacher working alone, it is very difficult to maintain correct exam writing procedures. In reality it is better to share exam writing duties with colleagues. By organizing exam writing teams in which all teachers giving the same course divide up the material, by developing exams for different units, each individual teacher works less. Instead of developing eight different unit exams, a teacher could join four colleagues teaching the same course and just write two exams, pretest and revise them, print and distribute them and even analyze the results and further revise the exam for future use. As a result of this more formal organization of exam writing tasks, an exam file can be developed in which different versions or cycles of specific exams can be stored and in a few years (if the textbook or program is not changed) the number of new exams that need to be developed will be greatly reduced.

Sharing test writing duties can also give continuity to courses. In order to share exams, teachers have to teach similar material, the result of which is that students will learn at a similar rate and it will be easier to assume that students finishing a course will begin the following course with similar abilities. However, in order to do this, it is important to clearly define what is to be tested and to be consistent from one exam version to another. Each version or cycle of an exam should be based on <u>one</u> and <u>only one</u> analysis of the content of the course. Each teacher should not just write the test about what he did in class.

In a cooperative exam development program there is a need for some kind of system to follow to analyze material to be tested. Hughes says that the "fuller the information on content, the less arbitrary should be the subsequent decisions as to what to include in the writing of any version of the test." (Hughes 1989: 49) We must try to have the tests reflect what went on in class. Nevertheless, since not all teachers teach exactly the same way, it would be impossible to develop an exam that would reflect what each individual teacher did in class. Probably the best way to analyze what should be included on the test is to carefully examine the textbook or program used in the courses. While not all people teach alike, they do base what they will do in class on some program or model. If the exams are based on the content of the textbook or program, all teachers, besides doing whatever extra activities they usually include in their classes, are committing themselves to covering the material in the program. Therefore, all students are at least finishing the program together and the results from different versions of the exams will be more valid.

On analyzing the material to be covered on an examination, it is preferable to consistently use the same chart or grid for each course and unit. Figures 1 and 2 below can be adapted for most teaching situations.

Figure 1 is used to determine the relative weight of different structures (in the case of a grammar exam) or strategies (in the case of a reading or listening exam). This figure can be used to estimate the percentage of time spent on a given structure/strategy. The structure or strategy is written in the first column, the exercises or practices in the book which contain the structure are listed in the second column. These

practices are counted up and the number of practices is written in the third column. After all practices have been listed and counted, the total number of practices is written at the bottom of the grid (*Total*). This total can then be used to determine the percentage or relative weight of each structure in the Unit.[6] (Column Four.) A hypothetical example is presented in Figure 3.

Once the relative weight of each structure and/or strategy has been determined, the results are copied onto Columns One and Two of the Exam Planning Grid (Figure 2). These two columns are prepared once before the first version or cycle of an exam is written and are used for all further exams of the same material. Columns Three, Four and Five can vary from one exam version to another as the exam writer changes formats and subtly varies the composition and size of the sections.

In Column Three the exam writer decides what format will be used to test the structure or strategy. The hypothetical case presented in Figure 3 shows how one structure can be tested with more than one exam section (*Past tense: Answers and Fill in*) or two or more structures can be joined to create a more integrative section (*Affirmative (+) and negative (-) statements tested together*). Also with some minor changes the grid can be adapted for integrative tests. The formats listed in Column Three should reflect formats used in the textbook or that are used by all teachers. An examination situation is not the right time to present the students with formats they have never seen before.

---

[6] Calculate percentage by dividing each number of practices in Column 3 by the total number of practices. This number is a decimal. Convert it to a percentage by multiplying by 100.

The number of items included in each section is written in Column Four and in Column Five the number of points per item is included. Column 6 lists the total number of points for the section. The product of Column Four and Column Five should be similar to the percent in Column Two and the total of the products should be 100 (if it is a 100 point test.) Notice that the percents have been slightly modified in the final grid, but they are similar to the analyzed percentages.

A grid of this type is tedious to elaborate the first time. But, once the material has been analyzed, it does not have to be reanalyzed for future versions of the examination. The use of the same analysis for all tests of the same material will increase the validity of exams across cycles and lead to more consistent testing and grading of students.

References

Henning, Grant. 1987. A Guide to Language Testing: Development, Evaluation, Research. Cambridge, Newbury House Publishers.

Hughes, Arthur. 1989 Testing for Language Teachers. Cambridge University Press.

Figure 1
Grid to Determine Weight of Different Structures or Strategies in a Given Unit

Course _____ Textbook: _____ Unit _____

Column: 1	2	3	4
Structure/strategy	Book Practices	Number	Percent
		Total:	100%

Figure 1. A grid that can be used to analyze how many times a structure or strategy occurs in a given unit.

Figure 2
Exam Planning Grid

Columns: 1	2	3	4	5	6
Structure/strategy	Percent	Format	Number	Points	Total

Figure 2. A grid that can be used to plan format and organization for a given exam. Columns One and Two are prepared once, before the first version or cycle of an exam is written and are used for all further exams. Columns Three, Four, Five and Six can vary from one exam version to another.

Figure 3
Grid to Determine Weight of Different Structures
or
Strategies in a Given Unit
(Hypothetical Example)

Course *3* Textbook: *English 3* Unit *4*

Column: 1	2	3	4
Structure/strategy	Book Practices	Number	Percent
Past tense:			
+ statement	*3, 4, 5, 6, 8, 10*	*6*	*30%*
- statement	*5, 6, 8, 9, 10*	*5*	*25%*
question	*7, 8, 9, 10*	*4*	*20%*
Complement pronouns	*1, 2, 3, 9,10*	*5*	*25%*
	Total:		100%

Exam Planning Grid

Columns: 1	2	3	4	5	6
Structure/strategy	Percent	Format	Number	Points	Total
Past tense:		*Answer ques.*	*8*	*3 each*	*24 pts.*
+ statements	*30%*	*Fill in*	*11*	*3 each*	*33 pts.*
- statement	*25%* 57%				
question	*20%-21%*	*Give ans. Write ques*	*7*	*3 each*	*21 pts.*
Complement pronouns	*25%-22%*	*Fill in*	*11*	*2 each*	*22 pts.*

Figure 3. A hypothetical example.