DIFFERENTIAL ITEM DIFFICULTY IN TEST DEVELOPMENT AT EDUCATIONAL TESTING SERVICE

Purpose. Measures of differential item difficulty (DIF) can help test developers to identify questions that may be unfair for members of certain groups. DIF can, therefore, be an extremely useful tool in test construction and analyses. The purpose of this paper is to describe how measures of differential item difficulty have been integrated into the test development process at Educational Testing Service.

Overview. This paper will explain the meaning of DIF, briefly describe the generation of a useful DIF statistic, and discuss how DIF is applied at several major stages of the test development process.

Meaning of DIF. If a test question is harder for members of one group than for members of some other group, then some aspect of the question could be making it unfair, or "biased." However, there could be a real difference between the groups in knowledge of what the question is measuring. How can we tell if a question is accurately reflecting real differences or if the question itself is somehow producing unfair differences?

As one approach to the problem, we begin with the very reasonable assumption that if people know the same amount about what is being tested, then they should perform in much the same way on a test question regardless of differences in sex, race or ethnicity. If we could match people in terms of relevant knowledge and skill, then people in the matched groups should perform in similar ways on individual test questions.

In operational use of indices of differential item difficulty, people are matched on the basis of test scores or subscores. The scores can be shown to be reliable and valid, and they are obtained under the same conditions for all examinees. Even though people with the same test scores are not identical, they are likely to be reasonably well matched in terms of the knowledge and skill measured by the test.

Differential item difficulty occurs when people of approximately equal knowledge and skill perform in substantially different ways on a test question. Indices of DIF thus help to identify differences in difficulty caused by characteristics of the question itself, because real differences in relevant knowledge and skill have been accounted for to the extent allowed by the matching process.

It is important to realize that DIF is not a synonym for "bias." Professional judgment is required to determine whether or not the difference in difficulty shown by the DIF index is unfairly related to group membership. Black examinees, for example, may find a question about Harriet Tubman to be easier than would a matched group of White examinees. The relationship of question difficulty to group membership would be clear. Whether or not the relationship is unfair would depend on what the test was supposed to be measuring. If knowledge of that aspect of American history had been a valid element of the test specifications, then judges might consider the question to

Michael Zieky, Educational Testing Service, April, 1988

be fair in spite of its relationship to group membership. If, however, measurement of that type of knowledge were not supported by the test specifications, judges would consider the question to be unfair.

The fairness of a test question depends directly on the purpose for which a test is being used. For example, a science question that is differentially difficult for women may be judged to be fair in a test designed for the certification of science teachers because the question measures a topic that every entry-level science teacher should know. However, that same question, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry-level teachers. Appropriate use of DIF, therefore, requires that procedures be developed that incorporate the judgments of trained test developers and subject-matter specialists.

The Mantel-Haenszel Statistic. The DIF measure in use at ETS is based on the Mantel-Haenszel statistic which was first applied in medical research. The statistic was used to investigate such issues as the odds that smokers would develop cancer in comparison to the odds that a matched group of non-smokers would do so. In its use with tests, the Mantel-Haenszel statistic is based on a comparison of the odds of answering a question correctly for matched people in the groups being compared.

If, for example, a test has 65 questions, the people who have taken the test can be divided into as many as 66 clusters based on their test scores: one cluster containing people with scores of zero, another cluster containing people with scores of one, and so on up to a final cluster containing people with scores of 65. Even though no test can be a perfect measure of any knowledge or skill, the people within each cluster should be quite similar in terms of what the test is measuring.

The procedure looks within each cluster and calculates the odds that members of the two groups being compared will answer the question correctly. For example, if there are 20 women and 16 of them answer correctly, then the odds are 16/4 or 4 to 1 that a woman in the cluster will answer correctly. If 12 out of 18 men answer the question correctly, then the odds are 12/6 or 2 to 1 that a man in the cluster will answer the question correctly.

The next step in the procedure is to calculate the ratio of the two odds to obtain an indication of the relative advantage of one group over the other within the cluster. For our example, the ratio is 4/1 (the women's odds) divided by 2/1 (the men's odds), which equals 2. This indicates that women are twice as likely as men within the cluster to answer the question correctly. The "odds ratios" are then averaged across all of the clusters. (People who would like to know more about the statistic should read: Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.)

The Mantel-Haenszel statistic can be defined as the average factor by which the odds that members of one group will answer a question correctly exceed the corresponding odds for comparable members of the other group.

The Mantel-Haenszel statistic is, therefore, in the form of an odds-ratio. Even though odds-ratios have certain desirable statistical properties, the numbers are not intuitively meaningful for most people. (Is an odds ratio of 1.23 big or small?)

A Test Development DIF Statistic. To obtain a statistic that is more meaningful to test developers, the odds-ratios have been transformed to an index that can be interpreted directly in terms of differences in the difficulty of questions. The DIF statistic is expressed as differences on the delta scale which is commonly used by test developers at ETS to indicate the difficulty of test questions. For that statistic, known as MH D-DIF, a value of 1.00 means that one of the two groups being analyzed found the question to be one delta point harder than did comparable members of the other group. (Except for hard or easy questions, a difference of one delta is approximately equal to a difference of 10 points in percent correct between groups.) We have adopted the convention of having negative values of MH D-DIF mean that the question is differentially more difficult for the "focal group" (generally Asians, Blacks, Hispanics, Native Americans or females). Positive values of MH D-DIF mean that the question is differentially more difficult for the "reference group" (generally whites or males). Both positive and negative values of the DIF statistic are found and are taken into account by our procedures.

Categories of Questions. On the basis of MH D-DIF, questions are classified into three categories. The category into which a question will be placed depends on two factors: the absolute value of MH D-DIF and whether or not the value is statistically significant. "Absolute value" means without regard to the sign of the number. The absolute value of a number depends on how far away the number is from zero and not on whether the distance from zero is in a positive or in a negative direction. For example, .75 and -.75 have the same absolute value. If a value is "statistically significant" it is not likely to have been caused by chance alone. Statistical significance is not the same as practical importance. Even an extremely small DIF value could be "statistically significant" merely because the analysis had included a large number of examinees.

The three categories carry the labels A, B and C. Category A contains the questions with negligible or non-significant DIF. Categories B and C contain questions with statistically significant values of DIF. Category B contains the questions with slight to moderate values of DIF, and Category C contains the questions with moderate to large values of DIF. The procedures for using DIF described below are based on the categories into which the questions have been classified.

DIF in Setting Specifications. One of the most important stages of the entire test development process is that of setting specifications--establishing the blueprint that will guide the rest of the developmental effort. Test specifications detail the content and skills that are to be measured and indicate how those attributes are to be measured. The setting of specifications is most often done by committees of subject-matter experts, instructors, and job incumbents all of whom may use data from curriculum surveys, task analyses of behavior on the job, inspection of widely used texts, mail surveys, and interviews.

As more DIF data are gathered over time and as more research is performed, certain reasonably consistent relationships between aspects of the contents of test questions and the size and direction of their DIF values are becoming known. These relationships will be made available to the committee members as they determine the content and skills to be measured. At ETS a "feedback loop" has been established between the DIF data gathering and analysis activities and the establishment of specifications for new or revised tests. For each program that gathers DIF data, an experienced test developer has been given the responsibility of preparing summary reports for use by the committees.

A Note on Sensitivity Review. Every question in every test developed at ETS is scrutinized by specially trained reviewers who follow an extensive set of guidelines to ensure that questions are not offensive, do not reinforce negative stereotypes, and that the questions reflect our multicultural society. It is important to note that DIF is not a replacement for that sort of scrutiny. Even if a question shows no differential difficulty it should not be included in a test if it fails the sensitivity review criteria. The appropriate use of DIF data is as an additional safeguard to help ensure the fairness of test questions. DIF should not be used as an excuse to eliminate the necessary reviews of questions.

DIF in Test Assembly. Once questions have been written and reviewed, they may be administered to students in a pretest to gather data before the questions are scored in a final form. Ideally, a large number of pretested questions will be available in a "pool" from which the test assembler will select the best set of questions to meet the specifications that have been set for a final form.

To ensure that DIF will be used appropriately in building final forms from pretested questions, the following procedures have been instituted:

- The content and statistical specifications for the test must be met.
- Large form-to-form variations in DIF in tests made from the same pool should be avoided. Test assemblers making more than one test from a pool of questions should not use up all of the questions in Category A in the first test to be assembled, thereby forcing later tests to have progressively larger DIF values.
- Within the above constraints, questions from Category A should be selected in preference to questions from Categories B or C.
- For questions in Category B, when there is a choice among otherwise equally appropriate questions, then questions with smaller absolute DIF values should be selected in preference to questions with larger values.
- Questions from Category C will NOT be used unless they are judged to be essential to meet test specifications.
- If Category C questions must be used, the test assembler will document the reason why and a reviewer will check to make sure that the use of Category C questions was indeed necessary.

The procedures have been designed to result in the selection of the most appropriate set of questions in a final form that meets all of the content and statistical specifications that have been established.

DIF Before Score Reporting. Not all testing programs are able to pretest questions before they are used in a final form. Such programs must still apply DIF if sufficient samples of examinees in the various groups can be obtained during the brief period after the test has been administered, but before the scores have been released. The analyses performed at that stage allow the identification and removal of questions that have been judged to be unfairly related to group membership before those questions affect the scores of any examinees. To avoid the possibility that test developers may be too lenient in judging questions which they have previously selected for inclusion, the procedures established at ETS require that the questions identified at that stage pass multiple reviews by people who had not worked on the test before, if the questions are to be retained for scoring.

DIF After Score Reporting. Once test scores have been released, further DIF analyses may be performed if sufficient additional samples of people are available to provide more information than was previously obtained. These post-hoc analyses provide the most stable data because they are performed on the largest samples. The analyses are rich sources of data for the generation and confirmation of hypotheses about the causes of DIF.

Conclusion. In all of our work with the index of differential item difficulty, as with any other statistic, we must constantly keep in mind the fact that numbers cannot make decisions for us. The subject of differential item difficulty is extremely complicated and it touches on some very sensitive issues. We must take great care not to allow some quantitative system to take the place of informed judgment. We have an obligation to test-takers, test-users, and to the public to ensure to the best of our ability that tests developed by ETS are free of questions that are inappropriately difficult for different groups.