# A

# DIF Primer

Michael Zieky
Educational testing Service
2003

Introduction. Measures of differential item functioning (DIF) are used at Educational Testing Service (ETS) to help ensure the fairness of tests. The process of using DIF relies on both statistical analyses and human judgments about test questions.

This paper explains what DIF is and describes how ETS test developers use DIF information to make tests as fair as possible. The paper has been written for people who do not have specialized knowledge about measurement. References are provided so that readers can obtain additional information, if they wish.

What is DIF? In addition to reviews of test questions to help ensure their fairness[1], ETS uses an empirical measure of differential item functioning based on the actual test performance of examinees in different groups.

If people in different groups perform in different ways on a test question, the groups could really differ on the relevant knowledge or skill being tested, or the question could be unfairly causing a difference to appear.[2] How can we tell whether the difference is real or if some aspect of the question itself is causing the difference to appear?

If people in different groups could be matched in terms of relevant knowledge and skill, then people in the matched groups could generally be expected to perform in similar ways on test questions.

[1] For more information about the reviews for fairness, see ETS Fairness Review Guidelines. (2003), Princeton, New Jersey: Educational Testing Service.

[2] For a discussion of how the fairness of test questions is judged, see Zieky, M. J. (1992). Evaluating Items for Fairness. CLEAR Exam Review. III, 2. For a general discussion of fairness issues, see R. A. Berk (Ed.), (1982) Handbook of methods for detecting test bias, Baltimore: Johns Hopkins University Press. See also Camilli, G. & Shepard, L. A. (1994) Methods for identifying biased test items. Thousand Oaks, CA: Sage.

**Differential item functioning (DIF) occurs when people of approximately equal knowledge and skill in different groups perform in substantially different ways on a test question.**[3] Measures of DIF thus help to identify questions that may be unfair because group differences in relevant knowledge and skill have been taken into account to the extent allowed by the matching process.

<u>How is the matching done?</u> The common approach taken at ETS and elsewhere is to use scores on a test or test section as the means of matching people on the knowledge and skills measured by that test or section.

Test scores are not perfect criteria for matching because no test can be perfect. However, the scores are derived from tests carefully designed to measure relevant knowledge and skill. The validity of the scores is empirically assessed, they are shown to be reliable (consistent), and they are obtained under comparable conditions for all examinees. These characteristics usually make test scores preferable to any other generally available options as a way of matching people.[4] People who score the same number of questions correct on a test may not be identical, but they are certainly reasonably closely matched on whatever knowledge and skill the test is measuring.

Even though test scores are almost always the best matching criteria available, there is concern about the perception of the circularity of using potentially biased test scores as a criterion to search for biased questions. That problem is addressed during the process of validation by gathering various kinds of evidence to demonstrate that the inferences made on the basis of the test scores are appropriate. The more evidence there is that the test scores allow appropriate inferences, the more comfortable we can be that the scores are reasonable matching criteria.

At ETS, a further step is taken to help ensure that the test scores used for matching are themselves free of questions that may be unfair. First, a preliminary DIF analysis is done and any questions with elevated values of the DIF statistic are removed from the test scores on which people are to be matched. The modified test scores are then used to match people for the calculation of the DIF statistics that are to be used operationally.

[3] For a detailed discussion of DIF including history, methodology, practical uses, and empirical findings see P. Holland and H. Wainer, (1993) <u>Differential item functioning,</u> Hillsdale, NJ: Lawrence Erlbaum.

[4] For a survey of the range of technical approaches to comparing items across groups (all of which use some procedure to match groups), see N. S. Cole and P. A. Moss, P.A., Bias in test use, in R. L. Linn (Ed.), (1989) <u>Educational measurement</u>, Washington, DC: American Council on Education.

Is DIF the same as bias? No statistic can determine whether or not a test question is biased. DIF helps us spot questions that may be unfair, but DIF is NOT synonymous with bias.

One reason DIF is not proof of bias is that the matching process is necessarily imperfect. Tests usually measure more than a single aspect of knowledge or skill. A perfectly fair question may show DIF if it happens to be measuring a skill that is not well-represented in the test as a whole. For example, if a mathematics test has many algebra questions and only a few geometry questions, matching people on total scores will primarily match them on algebra knowledge and may not match them well in terms of geometry knowledge. The geometry questions may show high values of DIF simply because the groups were not matched well on geometry knowledge, not because there is anything unfair about the questions.

DIF may also occur if a topic is of greater interest to some group(s) than to others, or if members of some group(s) are more likely to be exposed to the information being tested. In those cases, judgment is required to determine whether or not the difference in difficulty shown by the DIF index is unfairly related to group membership. For a question to be considered unfair, the DIF must be caused by some aspect of the question that is NOT related to what the question is intended to measure. The fairness of a question depends on the purpose of the question. A question may be fair for one purpose and unfair for a different purpose.

For example, women taking a licensing test for nurses may find a question concerning breast cancer easier than do a matched sample of men. If the question measures information that all nurses ought to know, the question would be fair in spite of the difference. The same question, however, might be considered unfair on a test of general knowledge taken by people without specialized training in nursing.

What is the Mantel-Haenszel Statistic? One of the DIF measures in use at ETS and elsewhere is based on the Mantel-Haenszel (M-H) statistic.[5] At ETS, the M-H statistic is expressed on a scale in which negative values indicate that the question is more difficult for members of the "focal group" (generally African American, Asian American, Hispanic American, Native American, or female examinees) than for matched members of the "reference group" (generally White or male examinees). Positive values mean that the question is more difficult for members of the reference group than for matched members of the focal group.[6]

[5] Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease, Journal of the National Cancer Institute, Vol. 22, p. 719-748.

[6] For information about the statistical aspects of DIF, see:

Holland, P. & Thayer, D. Differential item performance and the Mantel-Haenszel procedure, in H. Wainer and H. Braun (Eds.), (1988) Test Validity, Hillsdale, NJ: Lawrence Erlbaum.

How is DIF applied to questions? On the basis of the M-H statistic, questions are classified into three categories: A, B, or C. The category into which a question will be placed depends on two factors: the absolute value of the statistic[7] and whether or not the value is statistically significant[8].

Category A contains the questions with little or no difference between the two matched groups. Category B contains questions with small to moderate differences, and Category C contains the questions with the greatest differences. The procedures described below for using DIF in test development are based on the categories into which the questions have been classified.[9]

How is DIF used in writing questions? Years of collected data on questions suggest that certain topics and contexts tend to be associated with higher than chance occurrences of Category C DIF. When sufficient evidence exists, test developers are told not to write such questions unless they are required for the measurement of some particular subject. For example, questions concerning military issues tend to be more difficult for women than for a matched group of men and would not be written for a general skills test. An American history test, however, would still require questions about America's involvement in wars.

How is DIF used in assembling tests? Once questions have been written and reviewed, they are, for many testing programs, administered to students in a pretest. (A pretest is used to try-out the questions rather than to generate scores for examinees.) Whenever pretest samples are large enough to include sufficient people in the various focal groups, DIF statistics are computed.

The primary goal of test assembly, whether DIF data are present or not, is to meet the required test specifications (the plan or blueprint for the test) with questions of high quality. When DIF statistics are available, test developers at ETS select questions from category A in preference to questions from other categories.

Dorans, N. (1989) Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method, Applied Measurement in Education, Vol. 2, No. 3, pp. 217-233.

[7] "Absolute value" means without regard to the sign of the number.

[8] If a value is "statistically significant," it is not likely to have been caused by chance alone.

[9] The DIF categories are based on differences expressed on the delta scale of item difficulty, known as MH D-DIF. Category A consists of items with MH D-DIF not significantly different from zero, or less than 1.0 in absolute value. Category C consists of items with MH D-DIF significantly greater than 1.0 and absolute value of 1.5 or greater. Category B consists of all other items.

However, the pool of category A questions may not be large enough to allow the test assembler to meet all of the test specifications for all of the editions of the test to be assembled from the pool. In that case, category B questions may be used. If there is a choice among otherwise equally appropriate questions in category B, and the parallelism of all the editions of the test to be assembled can be maintained, preference is to be given to the questions with smaller DIF values.

Test assemblers at ETS may not select any questions from category C unless the questions are essential to meet important test specifications and the factors that seem to account for the high level of DIF are judged not to represent bias. The selection of a category C question for a test requires documentation of the reason why such a question had to be used, independent corroboration by a reviewer that the selection was necessary, and justification that performance on the question is not unfairly related to group membership.

The rules for assembling tests using DIF statistics must be followed within the context of the need to produce editions of the test that are parallel to one another. Because more than one edition of a test is to be assembled from a pool of questions, the tests assembled earlier in the process should not use all of the Category A questions or all of the Category B questions with smaller DIF values. To do so would force tests assembled later in the process to have progressively increasing levels of DIF.

How is DIF used after operational administration? We are not always able to obtain a sufficient number of examinees at the pretest stage of test development to allow the computation of DIF. Furthermore, not all testing programs are able to pretest. Therefore, DIF statistics may be computed during the brief period after the test has been administered but before the scores have been reported if there are a sufficient number of examinees in the various focal groups. At this stage, examinees have already taken all the questions, and the issue is whether or not questions with elevated DIF values should be included in the test score.

When the analyses are performed at this stage, test development procedures require the review of all category C questions to determine whether or not the questions are unfairly related to group membership. If the questions are judged to be possibly unfair they must be removed from the scored portion of the test before those questions affect the reported scores of any examinees.

To avoid the possibility that test developers may be too lenient in retaining and defending questions they have previously selected, ETS procedures require that the category C questions identified at this stage pass multiple reviews by people who did not work on the test before. As an additional safeguard, external reviewers who are not affiliated with ETS participate in review sessions.

Conclusion. Because no statistic can prove that a test question is either fair or unfair, DIF is only one of the means ETS uses to help ensure our tests are as fair as possible such as  equal treatment of test takers, diverse external input to test content, fairness review, promotion of proper test use, and research on fairness issues.