DOCUMENT RESUME

ED 314 464 TM 014 339

AUTHOR Millsap, Roger E.; Meredith, William
TITLE The Detection of DIF: Why There Is No Free Lunch.
PUB DATE Jul 89
NOTE 12p.; Paper presented at the Annual Meeting of the Psychometric Society (Los Angeles, CA, July 7-9, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Equations (Mathematics); *Mathematical Models; *Scores; *Statistical Analysis
IDENTIFIERS *Conditional Observed Scores; *Differential Item Performance; Invariance Principle

ABSTRACT

Conditional observed score (COS) and latent trait (LT) definitions of differential item functioning (DIF) are explored to determine when they are equivalent. COS methods rely solely on observed measurements, and LT methods model the response to an item as a function of an unobserved hypothetical latent ability or trait. For the case of dichotomous test items, the COS approach defines DIF by population differences in the conditional probabilities of responding correctly to the item, conditioning on the observed ability measure. The LT approach defines DIF by population differences in the conditional probabilities of responding correctly to the item, conditioning on the latent trait. Although DIF detection methods are usually applied to dichotomous ability or achievement test items, the discussion focuses on a more general level, and the results may have applications for general questions of measurement invariance in multiple populations. It is concluded that the conditions under which COS and LT definitions are equivalent are quite specialized. Equivalence generally requires invariance of the conditional densities. Other conditions of equivalence are discussed. It appears that precise invariance of prior densities can rarely be assumed in practice. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314464

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it
Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Roger E. Millsap
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

The Detection of DIF: Why There is No Free Lunch.

Roger E. Millsap

Baruch College, The City University of New York

William Meredith

University of California-Berkeley

Paper presented at the annual meeting of the Psychometric Society, University of California-Los Angeles, July 6-9, 1989.

TM014339

BEST COPY AVAILABLE

Methods for detecting differential item functioning (DIF) across multiple examinee populations have been under development for several decades. We can classify the available methods in two categories. The first category includes methods that rely solely on observed measurements (item scores, subtest scores, or external measurements) for the detection of DIF. This category includes methods using transformed item difficulties (Angoff, 1982; Angoff & Ford, 1973) and also methods that examine the conditional association between item scores and population membership within ability groups defined by an observed measure (Holland & Thayer, 1986; Ironson, 1982; Marascuilo & Slaughter, 1981; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981). We will focus on these conditional methods, which will be denoted "conditional observed score" (COS) methods. The second category includes methods which model the response to an item as a function of an unobserved, hypothetical latent ability or trait. Within the hypothesized model, DIF is defined by population differences in the function relating the item response to the latent trait (Lord, 1980). Methods within this category will be denoted "latent trait" (LT) methods.

COS and LT methods rest on different definitions of DIF. For the case of dichotomous test items, the COS approach defines DIF by population differences in the conditional probabilities of responding correctly to the item, conditioning on the observed ability measure. The LT approach defines DIF by population differences in the conditional probabilities of responding correctly to the item, conditioning on the latent trait. When can these two definitions be considered equivalent

in theoretical terms? If the definitions are equivalent, the COS approach has an advantage in eliminating the need for computer-intensive model-fitting and estimation. If the methods are not equivalent, investigators may reach different conclusions regarding DIF depending on the method chosen.

In the following, we explore the conditions under which the COS and LT definitions of DIF will be equivalent. Although DIF detection methods are usually applied to dichotomous ability or achievement test items, the discussion will be at a more general level, and the results may have applications for general questions of measurement invariance in multiple populations.

Conditions for Equivalence

Let u, v, and w be three random variables, possibly vector-valued. We will assume w to be a latent variable of interest, and u to be an observable variable that is intended to measure w. Concern for DIF will center on u. A second observable variable v will be used in studying the possible DIF in u through COS methods. The variables u, v, and w may be discrete or continuous, but our notation will treat all variables as continuous except where needed in discrete examples.

Let $h_i(w)$ be the density function for w (the prior density) defined with respect to the ith population of interest, i=1..S. These populations will ordinarily be defined in terms of observed variables such as ethnicity, gender, or age. Let $g_i(u|w)$ be the conditional density function for u given w in the ith population. We can also define the conditional density functions $f_i(u|v)$, $t_i(v|w)$,

$q_1(v;u,w)$, and $c_1(u;v,w)$. Finally, let $d_1(u,v;w)$ be the conditional joint density function for u and v given w.

We define measurement invariance or lack of DIF to hold for u as a measure of w when

$$g_1(u;w) = g(u;w) \qquad (1$$

for i=1..S. If u is a dichotomous test item, this definition is identical to the LT definition of DIF if the notation is altered to reflect the discrete nature of u:

$$g_1(u;w) = P_1(u=1;w) = P(u=1;w).$$

Invariance holds when the item characteristic curves are identical among the populations of interest (Lord, 1980). If u is continuous, we could define weaker forms of invariance than that given in (1). For example, if u is a vector of observed measures and w is a vector of factor scores within the common factor model, the usual definition of factorial invariance would involve only the conditional first and second moments of u (Meredith, 1964a, 1964b).

COS definitions of DIF employ the conditional density $f_1(u;v)$. We can define COS invariance for u with respect to v as

$$f_1(u;v) = f(u;v) \qquad (2$$

for i=1..S. If u is a dichotomous test item and v is the unweighted total test score (possibly omitting u), the definition in (2) reduces to the null hypothesis examined in most chi-square-based COS procedures if

$$f_1(u;v) = P_1(u=1;v) = P(u=1;v).$$

When will measurement invariance as defined in (1) be equivalent to invariance as defined in (2)? Is it possible to have invariance in

$g(u;w)$ but not in $f(u;v)$? Is the converse possible?

First, it can be easily shown that the two definitions need not be equivalent generally. Assume that invariance holds as in (1). Also assume local independence between u and v with respect to w. In other words, if $d_1(u,v;w)$ is the conditional joint density for u and v, we have

$$d_1(u, v;w) = g(u;w)\ t_1(v;w). \qquad (3$$

Then we can express $f_1(u;v)$ as

$$f_1(u;v) = \frac{\int_w g(u;w)\ t_1(v;w)\ h_1(w)\ dw}{\int_w t_1(v;w)\ h_1(w)\ dw} \qquad (4$$

From this equation, it is clear that (2) will generally hold only if the product $t_1(v;w)\ h_1(w)$ is invariant. In particular, population differences in the prior densities $h_1(w)$ can result in differences in the conditional densities $f_1(u;v)$.

A practical example of this sort would occur if u is a dichotomous item score in a test containing p items, v is the sum of the remaining p-1 item scores on the test, and all items follow the Rasch model with latent trait w. In this case, u and v are locally independent with respect to w. Invariance in 2) may not hold even if the condition in (1) does hold.

As this example illustrates, local independence between u and v with respect to w is an important consideration in evaluating the equivalence of (1) and (2). Equation 4 suggests that the two definitions need not be equivalent when u and v are locally independent. Clearly, special cases may exist in which the two definitions are

equivalent in spite of local independence between u and v. For example, if both the prior densities $h_1(w)$ and the conditional densities $t_1(v|w)$ are invariant, the definitions are equivalent.

Suppose that local independence does not hold for u and v. Then (4) can be written

$$f_1(u|v) = \frac{\int_w c_1(u|v,w)\ t_1(v|w)\ h_1(w)\ dw}{\int_w t_1(v|w)\ h_1(w)\ dw} \qquad (5$$

Now consider the special case in which v is a sufficient statistic for w. Then $c_1(u|v,w) = f_1(u|v)$, and this conditional density does not involve w. Population differences in the prior densities $h_1(w)$ will not prevent invariance in $f_1(u|v)$. Suppose that (1) also holds. Must (2) then hold in this case?

The answer is no, not in general. Given sufficiency, we know that

$$c_1(u|v,w) = f_1(u|v) = \frac{g(u|w)\ q_1(v|u,w)}{t_1(v|w)} \qquad (6$$

Note that we cannot have $q_1(v|u,w) = t_1(v|w)$ because this implies local independence of u and v. Then (2) will hold only if the ratios

$$\frac{q_1(v|u,w)}{t_1(v|w)}$$

are also invariant for i=1..S.

As an example, suppose that u is a dichotomous item score, v is the sum of p item scores including u, and all items follow a Rasch model with latent trait w. Then v is sufficient for w, but u and v are not locally independent. In this case,

$$q_1(v|u,w) = q_1(v_{p-1}|w),$$

where v_{p-1} is the sum of the p-1 items excluding u. Then the $f_1(u|v)$ are invariant if the ratios

$$\frac{q_1(v_{p-1}|w)}{t_1(v|w)}$$

are invariant. Since $g(u|w)$ is assumed invariant, the ratios are invariant if the numerators of the ratios are invariant, or (1) holds for the p-1 items excluding u.

In the above case, we have assumed that v is a sufficient statistic for w. Suppose that v is not sufficient for w. If local independence does not hold for u and v, when are definitions (1) and (2) equivalent? In this case, (5) does not reduce to any simple form in general. The definition in (2) generally holds only if the prior densities are invariant, and if $c_1(u|v,w)$ and $t_1(v|w)$ are both invariant. Again it is possible that for some specific choices of prior densities, $c_1(u|v,w)$ and $t_1(v|w)$, the two definitions can be made equivalent.

A familiar example of the general case occurs when u is a dichotomous item score, v is the unweighted sum of p item scores including u, and all items follow a two-parameter logistic model with latent trait w. Local independence does not hold between u and v, and v is not sufficient because it is an unweighted sum. Algebraically, it can be shown that if (1) holds, (2) generally holds only if (1) also holds for the p-1 items excluding u and if the prior densities are invariant.

The foregoing results show that when u and v are not locally independent, the sufficiency of v with respect to w is an important consideration in determining the equivalence of definitions (1) and (2). When sufficiency holds, equivalence does not require invariance of the prior densities $h_1(w)$. If v is not sufficient for w, population differences in these densities will generally prohibit the equivalence of (1) and (2).

Conclusion

The conditions under which the COS and LT definitions are equivalent are quite specialized. First, equivalence generally requires invariance of the conditional densities $t_1(v|w)$. In the COS approach, this entails careful selection of the observed measure v to avoid diffential functioning in this measure. This fact is generally recognized (Ironson, 1982). Secondly, the precise conditions for equivalence depend on both the local independence of u and v, and the possible sufficiency of v for the latent trait w. If u and v are locally independent, equivalence generally requires invariance in the prior densities and in the conditional densities $t_1(v|w)$. Note that local independence of u and v precludes sufficiency of v for w in the cases of interest. But if local independence does not hold, the sufficiency of v for w is important. Given sufficiency, the equivalence of (1) and (2) does not require invariance of the prior densities.

In practical applications, v is typically an unweighted sum of item scores. If local independence can be assumed among these items with respect to a latent trait w, local independence between u and v simply

depends on whether u is included in the summation leading to v. There is an advantage to including u in the summation for v, thereby removing the local independence. This point was demonstrated by Holland and Thayer (1986) in a slightly different context. On the other hand, an unweighted sum of item scores will be sufficient for w only when the items fit the Rasch model with latent trait w, or when all items have identical discrimination parameters. Hence considerations of sufficiency may have limited practical value.

If v is not sufficient for w, population differences in the prior densities $h_1(w)$ will generally prevent the equivalence of (1) and (2). Precise invariance of the prior densities can rarely be assumed in practice. Since sufficiency of v is also unusual, we must conclude that formal equivalence between (1) and (2) will be the exception, rather than the rule.

References

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.) Handbook of Methods for Detecting Test Bias, Baltimore, MD: The Johns Hopkins University.

Angoff, W.H. & Ford, S.F. (1973). Item-race interaction in a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Holland, P.W. & Thayer, D.T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Ironson, G.H. (1987). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.) Handbook of Methods for Detecting Test Bias, Baltimore, MD: The Johns Hopkins University

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems, Hillsdale, NJ: Erlbaum.

Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on x^2 statistics. Journal of Educational Measurement, 18, 229-248.

Meredith, W. (1964a). Notes on factorial invariance. Psychometrika, 29, 177-185.

Meredith, W. (1964b). Rotation to achieve factorial invariance. Psychometrika, 29, 187-206.

Scheuneman, J.D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.