tions, error variances, and so on, avoiding the technicalities needed for complex items. In the final section it would be minimally sufficient for the student to accept (14.20) as the expression for subtest convergent and discriminant validities, as illustrated in Table 14.6.

END NOTES

The material in this chapter is based on McDonald (1982b). A very similar account is in McDonald (1997a). Other approaches to multidimensional item response theory are to be found in the literature. A treatment based on direction vectors has been given by Reckase (1985). See Reckase (1997) for a relatively nontechnical account. It is my belief, as should be clear, that independent clusters provide the best basis for applications, but further work could change the picture.

1. These data are used by kind permission of the American College Testing program.
2. The NOHARM program, included in the set on diskette available from Lawrence Erlbaum Associates, was designed to fit multidimensional item response models with confirmatory or exploratory independent cluster structures.
3. See Reckase (1997).

In "Test Theory: A unified treatment" (1999).
R.P. McDonald, Lawrence Erlbaum.

Chapter 15

Comparing Populations

Throughout the earlier chapters the phrase *populations of interest* has been used to point to the fact that many of the indices and parameters dealt with are accidental parameters of the population sampled. Every respondent can be classified as belonging to an indefinite number of populations. There is that most obvious identification of individuals by gender. There are less obvious but commonly employed classifications (by self or imposed by others) in terms of "ethnic group"—those complex results of some thousands of years of cultural and political history marked by expansion, invasion, infiltration, conflict, and conquest across the surface of our planet that may give different citizens of a modern nation-state a chosen or imposed identity distinct from their citizenship. There are attempted classifications by "race," from the infamous imposed classifications that have marked segregationist societies to identities chosen by members of a disempowered group for whom racial pride may function positively in a movement toward empowerment. There are possible classifications on such cultural bases as religion, and classifications by socioeconomic status or educational level. Ever finer classifications can be obtained if we select groups by their score on one test to study their responses to others, until the individual is the intersection of so many properties that the individual is a population containing just one member.

Test theory, developed primarily in the context of cognitive tests, has been centrally motivated by educational applications. Because a major use of tests of abilities/aptitudes has been the selection of individuals for admission to college and university programs, and, in some countries, to selective high schools, the most obvious populations of interest are those whose existence has been given recognition in laws concerning discrimination. Although tests alone cannot be used to redress the wrongs created by discrimination and by the enormous inequalities of educational and economic opportunity that characterize most modern states, there is a plain duty on the part of test developers to ensure that the tests themselves are not sources of discrimination and inequity. Accordingly, there has been a considerable amount of effort in work on test theory to study the conditions under which a test score on individuals from distinct populations (a) measures the same attribute and (b) gives an unbiased estimate of the relative standing of the individuals on that attribute. Research applications of noncognitive tests—measures of values and attitudes, for example—also require a technology to check whether the test measures the same attribute in populations of interest to the investigation, and, if so, whether it gives an unbiased estimate of it. Thus, it is possible for men and women to perceive the same set of attitude items sufficiently differently to raise the question whether "the same" attribute is being measured in both genders. The main illustration in this chapter is a study of just this possibility.

In classical test theory, as we have seen, indexes such as reliability coefficients, and item parameters such as item difficulty measured by proportion passing, or item–test correlations, are incidental properties of the population sampled. It is considered to be an axiom of item response theory that, in contrast, the item parameters are invariant across populations of interest, provided that an appropriate common scale of the latent traits is adopted. This *axiom of invariance* is a mathematical tautology. It has sometimes been misunderstood to imply that if, say, a unidimensional item response model fits two populations of interest, the parameters must be the same in both populations (when the latent trait is measured on a common scale). What the axiom of invariance actually means is that if the item parameters from the two groups cannot be rescaled so as to coincide, we can always introduce further latent traits so that they do. This is trivially true, because we can always use population membership as a "latent trait"[1] and make a model whose parameters are tautologically invariant. In applications, it is not mathematically guaranteed—for a model with a fixed number of latent traits—that the item parameters can agree across populations. If they do not, we cannot strictly claim that the items measure the same attribute in these populations; what they have in common may be at least slightly different.

In a comparison of two populations, it has become an accepted convention to identify one population as the *reference group*, and the other as a *focal group*. "Reference group" is a synonym for what we have called the calibration group. As before, it determines the metric of the latent traits. In the sociopolitical context of the United States, for gender studies it is customary to choose males as the reference group; for education, it is

common to find comparisons of European Americans with African Americans, Hispanic Americans, Asian Americans, and/or Native Americans, with the European Americans as the reference group. It is perhaps futile to discuss here the possible biases that may determine such a choice, because a reference group generally must be chosen to determine the metric on which the item parameters are calibrated. It is enough to recognize that the choice is arbitrary, and is inconsequential from a measurement perspective.

In most treatments of this type of problem, initially each latent trait is scaled to be standardized—separately—in both reference and focal groups. If in fact the items will fit the same model with the same item parameters when the scale of the focal group is changed to standard score units taken from the reference group, then there are simple linear relationships between the sets of item parameters from which the change of scale can be determined. Corresponding approximate relationships will be revealed in estimates of these parameters from samples. If the two sets of item parameters cannot be made to agree by a change of scale, this fact will be revealed in departures of the item parameters from the expected linear relationships. *Coefficients of congruence* (agreement) between the sets of item parameters can be computed. As we show later, if the departures from a linear correspondence are sufficiently great and, accordingly, the coefficients of congruence are not large enough, we should not suppose that the items measure the same attribute in both groups.

In applications it may happen that a number of the items have parameters that are linearly related, and can be supposed to measure the same latent trait or traits, whereas some do not. It has become customary to say (somewhat redundantly) that if a binary item gives a different probability of its keyed response for subjects of the same ability in the reference and focal groups, the item shows *differential item functioning* (DIF). ("DIF" is more pronounceable than "DF"). For a more general definition of DIF, we say that a (quantitative or binary) item shows DIF if it gives a different mean response for examinees in different groups with the same value of the attribute (ability, attitude, personality trait, etc.). The concept of differential item functioning requires enough items to determine "the" attribute—that is, enough items that do not exhibit DIF—but this is probably not a very restrictive condition in applications. A necessary and sufficient condition for an item not to show DIF is that its item response function should be the same in the relevant populations. Recent research has provided a number of nonparametric devices intended to detect and evaluate DIF, using the score on a subset of the items as a substitute for the attribute.[2] Partly because these methods are rather technical, and have not yet been carefully evaluated, and partly because they do not easily fit our framework, we consider a direct method that applies classical factor-analytic concepts to the problem. This is the natural extension of the treatment of factor and item response models in previous chapters. The direct method also has a number of advantages that the nonparametric devices lack.

The next sections, accordingly, describe methods based on the classical treatment of factorial congruence that have the following properties:

1. They are applicable, in essentially the same way, to both quantitative and binary responses.

2. They can be applied equally to unidimensional and multidimensional data.

3. They provide a direct assessment of the amount of DIF in one or more items, and, more generally, of the agreement (congruence) of the parameters of all the items.

4. The analysis distinguishes three distinct types of DIF, namely, (a) differential item difficulty (known in the DIF literature as *uniform DIF*), (b) differential item discriminating power (referred to as *nonuniform DIF*), and (c) the effect of differential item dimensionality on an approximating model of lower dimensionality. (At the time of writing, researchers using item response models commonly fit unidimensional models to multidimensional data, although there is no good reason for this practice.)

5. Given the nature of the DIF, as in the previous point, we may hope to examine the item for the substantive cause of differential functioning.

6. The analysis yields an understandable assessment of the effect of one or more differentially functioning items in the set on the relative test score obtained when we exclude/include those items. That is, we can estimate the extent to which the test score may give a statistically biased estimate of the attribute, and therefore a judgment that is biased in a number of sociopolitical or legal senses.

7. The method provides an estimate of the mean and variance of the trait for the focal group in the metric of the reference group.

Taken together, these seven properties allow the test developer to make rational, substantively based decisions as to how to deal with the differentially functioning items, and how to develop further items.[3]

It is convenient to develop the procedure by example, carrying through a fairly detailed analysis of a paradigm case. In the next section we consider a unidimensional and a multidimensional linear model for quantitative responses, and the following section covers counterpart nonlinear unidimensional and multidimensional models for binary responses. A single data set is used to illustrate these developments. It is easily seen how these procedures may be applied more generally. For definiteness, and in line with the example to be used, we suppose that the populations of interest

are identified by gender, with males as the reference group and females as the focal group, and accordingly use m and f as identifying subscripts.

QUANTITATIVE RESPONSES

Suppose we have m items yielding quantitative item scores $X_1, \ldots, X_m$ from a random respondent. A likely source is (integer) Likert scores for ordered-category responses. We assume we may fit a linear common factor model to sampled values by normal theory, to a sufficiently good approximation. To identify the populations we attach a superscript m or f, writing $X_j^{(m)}$ and $X_j^{(f)}$ for the jth item score from the male and female populations, respectively. Suppose for the present that the items form a unidimensional/homogeneous set, fitting the simple Spearman single-factor model in each population. We write the model as

$$X_j^{(m)} = \mu^{(m)} + \lambda_j^{(m)} F_m + E_j^{(m)}, \tag{15.1a}$$

and

$$X_j^{(f)} = \mu^{(f)} + \lambda_j^{(f)} F_f + E_j^{(f)}. \tag{15.1b}$$

Here $\mu^{(m)}$ and $\mu^{(f)}$ are the item means in each group—counterparts of item "difficulty"—and $\lambda_j^{(m)}$ and $\lambda_j^{(f)}$ are item factor loadings—counterparts of item "discrimination," whereas F_m and F_f are the common factor/latent trait in each population, and $E_j^{(m)}$ and $E_j^{(f)}$ is the unique part of each item response, corresponding to the idiosyncratic property of the item. As before, we assume that the unique parts of the item responses are mutually uncorrelated.

The linear item response functions are the regression functions, given by

$$\mathcal{E}\{X_j^{(m)} | F_m = f_m\} = \mu^{(m)} + \lambda_j^{(m)} f_m, \tag{15.2a}$$

and

$$\mathcal{E}\{X_j^{(f)} | F_f = f_f\} = \mu^{(f)} + \lambda_j^{(f)} f_f. \tag{15.2b}$$

These functions are separately identified if we fix the scale by fixing the mean and variance of F_m and of F_f in their respective populations. Ordinarily we standardize both, setting means to zero and variances to unity.

Consider the items of the Illinois Rape Myth Acceptance Scale, listed in Table 15.1.[4] The responses are on a 7-point Likert scale, from *strongly*

TABLE 15.1
Illinois Rape Myth Acceptance Scale (Items Reordered)

1. When women talk and act sexy, they are inviting rape.
2. When a woman is raped, she usually did something careless to put herself in that situation.
3. Any woman who teases a man sexually and doesn't finish what she started realistically deserves anything she gets.
4. Many rapes happen because women lead men on.
6. In some rape cases, the woman actually wanted it to happen.
7. Even though the woman may call it rape, she probably enjoyed it.
10. When a woman allows petting to get to a certain point, she is implicitly agreeing to have sex.
11. If a woman is raped, often it's because she didn't say "no" clearly enough.
12. Women tend to exaggerate how rape affects them.
16. In any rape case one would have to question whether the victim is promiscuous or has a bad reputation.
18. Many so-called rape victims are actually women who had sex and "changed their minds" afterward.

5. Men don't usually intend to force sex on a woman, but sometimes they get too sexually carried away.
13. When men rape, it is because of their strong desire for sex.
14. It is just part of human nature for men to take sex from women who let their guard down.

8. If a woman doesn't physically fight back, you can't really say that it was a rape.
9. A rape probably didn't happen if the woman has no bruises or marks.
19. If a husband pays all the bills, he has a right to sex with his wife whenever he wants.

15. A rapist is more likely to be Black or Hispanic than White.
17. Rape mainly occurs on the "bad" side of town.

disagree = 1 to *strongly agree* = 7. The items have the character of beliefs that can be regarded as myths, in the sense that they may be widely held, but not on rational/evidential grounds, and they represent the cognitive/perceptual component of an attitude. It is possible[5] that acceptance of these statements serves distinct psychological functions for men and women—for the former, rationalizing/legitimating offensive behavior, and for the latter, denying vulnerability. Apart from conventional considerations, there is here an additional substantive reason for making the male population the reference group, namely, that a primary concern of research on rape myths is their specific predictive function for the behavior of males. The 19 items in the scale were selected to represent 19 subscales, each of five items, which had been very carefully constructed to reflect recognizably distinct facets of this false-belief/attitude complex. An examination of the item contents suggests a multidimensional structure, but for the first analysis we treat the data as unidimensional.

Data are available from $N = 368$ men and $N = 368$ women. Analyses in this and the next section employ the COSAN program. Table 15.2 gives means and variances of the item scores, and maximum likelihood (ML) factor loadings from the item covariance matrices. The Spearman model gives chi-squares (152 *df*) respectively of 344.05 for the male and 350.17 for the female samples, with RMSEAs both .059.

If the item parameters differ between groups only because of the choice of metric, then f_m and f_f in (15.2) are related by the scale transformation

$$f_f = kf_m + c, \tag{15.3a}$$

with inverse transformation

$$f_m = (1/k)f_f - (c/k). \tag{15.3b}$$

It then follows that

$$k\lambda_j^{(f)} = \lambda_j^{(m)}, \tag{15.4a}$$

and

$$\mu_j^{(f)} + c\lambda_j^{(f)} = \mu_j^{(m)}, \quad j = 1, \ldots, p. \tag{15.4b}$$

TABLE 15.2
Unidimensional Quantitative Responses

Item	$\mu^{(m)}$	σ_m^2	$\mu^{(f)}$	σ_f^2	λ_m	λ_f	μ_f^*	λ_f^*	$\mu_m - \mu_f^*$	$\lambda_m - \lambda_f^*$
1	2.88	3.01	1.87	1.87	1.13	0.88	2.96	1.10	−.08	.13
2	3.12	2.77	2.32	2.27	0.85	0.77	3.28	0.98	−.16	−.13
3	2.13	1.91	1.43	0.86	0.79	0.55	2.11	0.69	.02	.10
4	3.79	2.92	2.60	2.69	1.12	1.01	3.86	1.28	−.07	−.16
6	3.01	2.51	2.11	2.91	0.99	0.79	3.10	1.00	−.09	−.01
7	1.69	2.74	1.22	2.08	0.62	0.30	1.59	0.37	.10	.24
10	2.97	1.23	1.86	0.42	1.14	0.86	2.93	1.08	.04	.05
11	2.52	2.08	1.77	1.02	0.91	0.68	2.62	0.86	−.10	.05
12	2.25	1.15	1.53	0.33	0.88	0.58	2.26	0.74	−.01	.15
16	3.63	2.99	2.34	1.89	0.95	0.86	3.42	1.09	.21	−.14
18	3.40	2.32	2.50	1.58	1.03	0.89	3.61	1.12	−.21	−.09
5	4.24	2.28	3.47	1.37	0.67	0.74	4.39	0.93	−.15	−.26
13	3.91	3.52	2.79	3.21	0.85	0.60	3.54	0.76	.37	.09
14	2.39	2.35	1.89	1.81	0.76	0.44	2.44	0.55	−.05	.21
8	2.13	2.32	1.46	1.89	0.73	0.50	2.08	0.63	.05	.11
9	1.72	3.40	1.25	2.54	0.47	0.22	1.52	0.28	.20	.19
19	1.92	1.92	1.13	1.23	0.73	0.15	1.32	0.19	.60	.54
15	2.36	2.31	1.89	2.22	0.51	0.15	2.08	0.19	.28	.31
17	2.24	1.92	1.67	0.23	0.35	0.32	2.06	0.40	.18	−.05

By (15.4a), if the item parameters differ only because they are referred to the origin and unit determined by their own group, then in a graph of the loadings for the female group against the loadings for the male group, the points should lie on a straight line through the origin whose slope is k. Similarly, by (15.4b), in a graph of the differences in the means, $\mu_j^{(m)} - \mu_j^{(f)}$, against the loadings in the female group, the points should lie on a straight line through the origin whose slope is c. In graphs from sample data, there will be departures from the straight line due to sampling errors, and possibly departures due to differential functioning of some of the items. In the extreme, we might find a scatter of points about the lines suggesting actual overall failure of congruence—failure of agreement of at least a reasonable number of the item parameters. Without any more sophisticated technology, but with experience, we could identify differentially functioning items fairly successfully as points lying too far from a best-fitting straight line. As an exercise, the student is advised to plot these graphs, and make tentative judgments as to which items depart most from the expected straight line.

For a more careful procedure than mere inspection, we can carry out the following calculations: An estimate of the multiplier k from sample factor loadings given by

$$k = [\Sigma\, \lambda_j^{(m)}\lambda_j^{(f)}]/[\Sigma\, \lambda_j^{(f)2}] \tag{15.5}$$

minimizes the quantity

$$q_\lambda = \Sigma\, (\lambda_j^{(m)} - k\lambda_j^{(f)})^2 \tag{15.6}$$

and an estimate of the additive constant c given by

$$c = [\Sigma\, (\mu_j^{(m)} - \mu_j^{(f)})\lambda_j^{(f)}]/(\Sigma\, \lambda_j^{(f)2}) \tag{15.7}$$

minimizes the quantity

$$q_\mu = \Sigma\, (\mu_j^{(m)} - \mu_j^{(f)} - c\lambda_j^{(f)})^2. \tag{15.8}$$

This chooses a rescaling that makes the parameters as close as possible, when measured by a sum of squares of differences. The summation can be taken over all the items or over a subset believed to be free from DIF. They are easily if tediously computed by hand, or simple computer programs can be applied. In the example, the estimate of k is 1.261, and that of c is 1.246.

We note that relative to the zero mean and unit variance assigned to the male group as reference population, the mean and variance of the trait in the female group are, respectively, $-c/k$ and $1/k^2$. In our example, the mean of the latent trait in the female group is $-1.246/1.261= -.988$, and its variance is $1/k^2 = .629$. That is, the female group is both well below the male group on the average, and less diverse in their rape myth acceptance, in (male) standard score units. This is to be expected on substantive grounds.

The rescaled parameters from the female population, given by

$$\lambda_j^{(f)*} = k\lambda_j^{(f)} \tag{15.9a}$$

and

$$\mu_j^{(f)*} = \mu_j^{(f)} + c\lambda_j^{(f)}, \tag{15.9b}$$

may be compared to $\lambda_j^{(m)}$ and $\mu_j^{(m)}$, respectively. In our example, these and the resulting differences are given in the last four columns of Table 15.2. Simple inspection of the listed differences suggests that relative to the remainder of the items, item 19 has large differences in both loading and item mean, and the same is true, although less clearly, for item 15. Item 13 shows a notable difference in mean but not in loadings. Note also the positions of these items on the graphs.

The analysis gives standard errors for the item parameters, from which we can obtain approximate confidence bounds on the differences between them. We might regard confidence bounds that do not include zero as indicating significant DIF. But note that for a sufficiently small sample size, no item will show significant DIF, whereas for a sufficiently large sample size all items will; that is, no subset of items could be thought of as related by a scale change. The important question is the amount of the difference, not its technical "significance." The *SE*s of the loadings are all very close to .05 in both groups, and the *SE*s of the mean parameters are given by the item *SD*s divided by root sample size. These *SD*s, for the male group, range from 1.07 to 1.84, giving a mean *SE* on the order of .07. It should be an acceptable heuristic device to take a common set of approximate confidence bounds of $\pm.14$ $(= 2 \times .05 \times 2^{1/2})$ for the difference in loadings between the groups after rescaling, and $\pm.2$ $(\approx 2 \times .07 \times 2^{1/2})$, for the difference in means. From Table 15.2 we see that four items appear to have nonnegligible and "significant" differential slope parameters—in order, 19, with difference $.735 - .194 = .561$; 15, with difference .314; 7, with difference .244; and 5, with difference $-.226$. Three appear to have nonnegligible and "significant" differential mean parameters—in order, 19, with difference .60; 13, with difference .37; and 15, with difference .28—also possibly 16 and 18. (Tabulated values have been rounded.)

Burt's *coefficient of factorial congruence*

$$g_\lambda = (\Sigma\, \lambda_j^{(m)}\lambda_j^{(f)})/[(\Sigma\, \lambda_j^{(m)2}(\Sigma\, \lambda_j^{(f)2})]^{1/2} \qquad (15.10a)$$

(a "correlation coefficient not corrected for means") here measures the closeness of the loadings to agreement, without rescaling, and equals unity if and only if $q_\lambda = 0$ exactly. It is natural to define a corresponding coefficient of congruence for the difference in mean parameters $\mu_j^{(m)} - \mu_j^{(f)}$ by

$$g_\mu = [\Sigma\, (\mu_j^{(m)} - \mu_j^{(f)})\lambda_j^{(f)}]/[\{\Sigma\, (\mu_j^{(m)} - \mu_j^{(f)})^2\}\{\Sigma\, \lambda_j^{(f)2}]^{1/2}. \qquad (15.10b)$$

This measures the closeness of the item means to agreement, without rescaling, and equals unity if and only if $q_\mu = 0$. Again the summation can be over all the items or over a subset thought to be free of DIF. In our example, the congruence coefficients are $g_\lambda = .973$, $g_\mu = .968$, for all the items, and $g_\lambda = .991$, $g_\mu = .990$, with the suspect items, 19, 15, 13, 7, and 5, omitted.

If a number of the items are judged to have nonnegligible DIF, the effect of excluding/including these items can be assessed by comparing the relative test score characteristic curves for the retained items with that for the full set. Let $\{X_1, \ldots, X_r\}$ be any subset of the item scores. The relative test-score characteristic functions for the two groups are given in the metric of the male group, by

$$\mathcal{E}\{M_r \mid F_m = f_m\} = \mu_{rm} + \lambda_{rm} f_m \qquad (15.11a)$$

and

$$\mathcal{E}\{M_r \mid F_f = f_f\} = \mu_{rf}^* + \lambda_{rf}^* f_f \qquad (15.11b)$$

where M_r is the mean score on the selected items, μ_{rm} and λ_{rm} are means of the parameters for the male group over the r selected items, and μ_{rf}^*, λ_{rf}^* are means of the rescaled parameters for the female group. Noncoincidence of these functions is a precise specification of what we call (in the linear model) *differential test score functioning* (DTF). In our example we examine the effect of omitting versus retaining items 5, 7, 13, 15, and 19 on the relative test-score information functions. The expected values of the means of the item scores, as a function of the latent trait, for the full set, are

$$\mathcal{E}\{M_m \mid f\} = 2.753 + .815f$$

for the male group, and

$$\mathcal{E}\{M_m \mid f\} = 2.693 + .749f,$$

for the female group, whereas for the reduced set they are

$$\mathcal{E}\{M_r \mid f\} = 2.727 + .864f$$

for the male group, and

$$\mathcal{E}\{M_r \mid f\} = 2.732 + .840f$$

for the female group. In the range −3 to 3, the differences appear negligible. In sociopolitically significant applications it might nevertheless not be a sufficient reason to retain differentially functioning items that they have a negligible effect on the relationship between the test score and the attribute being measured. There is a cogent argument that justice should not only be done but should be clearly seen to be done.

We have seen that there is no mathematical reason why a unidimensional item response model—or model of fixed dimensionality—must have invariant item parameters. It is perfectly possible to find items with either loadings or means that cannot be brought into coincidence by rescaling the focal group to the units of the reference group. It might be conjectured, loosely speaking, that if an item shows DIF it must be because the item measures something in addition to the intended attribute in one of the populations but not the other.

Such conjectures require care in interpreting them. Using the methods of Chapters 9 and 14, it should be possible to determine whether in fact two populations require distinct models with different numbers of latent traits. If the mistake is made of fitting a unidimensional model to data that are in fact multidimensional, the parameters in the appropriate multidimensional models may be invariant with appropriate scaling, whereas the unidimensional approximation shows DIF. Thus DIF can consist of actual differences in loadings or item means in a unidimensional model or of apparent differences that result from the use of a unidimensional approximation to multidimensional data.

If the conjecture is that DIF is due to additional "dimensions" measured by the aberrant items, in general such "dimensions" cannot be interpreted strictly as latent traits. In the case of a single item, if we postulate that it measures "something in addition" to the recognized latent trait in one of the groups, the "something in addition" has the character of an item-specific component in that group. This is included in the item's unique component, and is not an additional dimension as ordinarily understood; that is, it is not a common factor/latent trait.

In the case of quantitative responses, a population-specific component produces a difference between populations in the unique variance of the item. But it will not induce, and hence cannot explain, a change in the slope parameter—the common factor loading. (In a standardized factor model, the unique variance is the unit complement of the communality and a change in unique variance changes the loadings, but not, as here, when the response is unstandardized.) If the conjecture is that there is a difference between populations in the mean of the unique component of the item, this is merely a tautological account of the irreducible difference—not removed by rescaling—in the mean parameter. The same is true for more than one item, if each is postulated to measure "something in addition" that is specific to itself. (There has been considerable confusion on this point in the literature on differential item functioning, on the part of researchers who are possibly not familiar with the formal properties of the common factor model.)

On the other hand, one way—although not the only one—in which differences in slopes could arise is indeed by fitting an approximating unidimensional model to data in which there is either a difference in dimensionality—in the number of latent traits—between groups, or distinct correlations between the traits. (And, to repeat, it still seems to be a common, although unfortunate practice, to approximate multidimensional data with a unidimensional model in applications of nonlinear item response models.) If, for example, items 19, 15, 13, 7, and 5 happen to define a separate factor in the female group, but not in the male group, or if there is such a factor in both groups, with a lower correlation for females between it and the factor defined by the complementary subset, this might account for the reduction in their loadings in the unidimensional approximation. It could also account for differences in the item means if these were related appropriately to the loadings. Such an analysis would give a nonvacuous account of DIF in terms of additional dimensions, that is, additional latent traits.

We note, however, that it would be inappropriate and usually ineffective to ignore substantive considerations, choosing to fit a two-dimensional structure with a second latent trait defined by the differentially functioning items. Thus, in the present example, items 19, 15, 13, 7, and 5 do not have mean differences proportional to their loadings. And they should, if they correspond to a second factor. Also, an examination of their item stems does not suggest that they share a distinct conceptual basis. It turns out that fitting this model (a) does not improve fit and (b) does not reduce the differences between parameters of these items.

When irreducible differences are found between item parameters, the possibility remains that (some) differences are due to fitting the unidimensional approximation to data whose substantive character requires a

multidimensional structure, and that the differences may vanish or at least change their pattern when an appropriate model is used. Besides, on general grounds we need methods for comparing populations and checking for DIF that apply to a multidimensional structure.

It is a very straightforward mathematical task to write down analogues of equations (15.1) through (15.11) for the case of two groups having r common factors, replacing (15.3) by a more general transformation. (It is also possible in principle to apply and slightly extend classical treatments of factorial congruence transformations to cases with unequal numbers of latent traits). However, such generality does not seem well motivated in the present applications. Rather, it is reasonable to restrict the transformation of metric to separate scale transformations, one for each trait. Consequently, we do not require a more general formulation, but simply apply (15.1) through (15.11) to each latent trait in turn. Note that such transformations will not alter the correlations between the traits.

A careful study of the nineteen item stems in Table 15.1 suggests the application of a more general model for the function of these myths, based on well-known mechanisms of blame, rationalization, and denial, to yield a four-factor model, namely:

I. Blame the victim: items 1, 2, 3, 4, 6, 7, 10, 11, 12, 16, 18.
II. Excuse the offender: items 5, 13, 14.
III. Deny it is an offense: items 8, 9, 19.
IV. Deny it happens "here" (the respondent's usual location): items 15, 17.

(Note that the items in Table 15.1 have been reordered in the tables to group as shown.) The results from fitting this model to the two groups are given in Table 15.3. Because the model contains no factorially complex items, it is convenient to write the parameters in single columns, rather than setting them out in the form of conventional 19×4 factor patterns. The four-factor model gives chi-squares (146 *df*) respectively of 241.17 and 296.32 for male and female groups, clearly fitting better than the unidimensional results. The improvement is greater in the male group. Applying (15.5) and (15.7) separately for each factor gives the scaling coefficients, coefficients of congruence, and relative means and variances of the trait in the female group shown in Table 15.4. It is of interest to note that the female group gives its smallest mean difference for factor IV, its smallest relative variance for factor III, and its largest mean difference and relative variance for factor II. These differences make substantive sense. Note also that all factor correlations are lower for the female group than the male group, suggesting clearer distinctions between these dimensions of rape myth acceptance, particularly between "blame the

TABLE 15.3
Multidimensional Quantitative Responses

	μ_m	μ_f^*	λ_m	λ_f	λ_f^*	$\mu_m - \mu_f^*$	$\lambda_m - \lambda_f^*$
1	2.88	2.92	1.14	.88	1.09	−.04	.05
2	3.12	3.24	.86	.77	.96	−.12	−.10
3	2.13	2.08	.79	.54	.68	.05	.11
4	3.79	3.82	1.14	1.02	1.27	−.03	−.13
6	3.01	3.06	1.00	.80	1.00	−.05	.00
7	1.69	1.57	.61	.30	.38	.12	.31
10	2.97	2.88	1.13	.86	1.06	.11	.07
11	2.52	2.58	.92	.68	.85	−.06	.07
12	2.25	2.22	.88	.58	.72	.03	.16
16	3.63	3.37	.96	.86	1.07	.26	−.11
18	3.40	3.57	1.05	.90	1.11	−.17	−.06
5	4.24	4.56	.73	.93	.95	−.32	−.22
13	3.91	3.61	.88	.77	.79	.30	.09
14	2.39	2.44	.80	.52	.53	−.05	.27
8	2.13	2.32	.94	.61	1.13	−.19	−.17
9	1.72	1.66	.63	.29	.54	.06	.09
19	1.92	1.38	.83	.18	.33	.54	.50
15	2.36	2.44	1.04	.46	.56	−.08	.48
17	2.24	2.73	.83	.89	1.08	−.49	−.25

Factor Correlations

$$\begin{bmatrix} 1. & .801 & .773 & .329 \\ .962 & 1. & .494 & .236 \\ .756 & .637 & 1. & .432 \\ .419 & .283 & .540 & 1. \end{bmatrix}$$

Note. In matrix, male results below the diagonal, female results above it.

TABLE 15.4
Scaling Constants and Female Means/Variances

	I	II	III	IV
k	1.240	1.026	1.849	1.211
c	1.188	1.064	1.399	0.718
$1/k^2$	0.650	0.950	0.293	0.682
$-c/k$	−0.958	−1.037	−0.757	−0.592

victim" and "excuse the offender," which for the male group appear to function as virtually the same concept.

However, we come next to the observation that on rescaling the loadings and mean parameters for each dimension separately, as in Table 15.3, we still appear to have nonnegligibly different loadings for items 5, 7, 15, and 19, and now also for item 14, and also different means for items 13 and 19, but no longer for item 15. The hypothesis that rape myth acceptance functions for men as a rationalization for offensive behavior and functions for women as a denial of vulnerability suggests the unidimensional analysis should indicate differences in slopes that are ultimately explained by distinct correlations of factor I, the most well-represented dimension, with factor IV, and possibly also with factors II and III. This does not seem to be the case. The functional type of explanation has an important general role in accounting for general levels and group differences in nonrational beliefs and attitudes, but a more fine-grained analysis would be needed generally to account for the behavior of individual items. Thus the notable behavior of item 15 fits a conception that by identifying rapists as "other" than themselves, men can avoid recognizing themselves as potential rapists, whereas women are less likely to identify rapists as "other" than the men they know. As another kind of "explanation," the even more notable behavior of item 19 might be "explained" by saying that for the male group this is integrated into a more general system of sexist beliefs that would not be shared by the female group. The other instances of DIF might similarly allow a specific account, rather than an account in terms of additional dimensions. Note in particular that the difference in mean for item 13 persists from the unidimensional analysis in Table 15.2. But in the context of the other two items, 5 and 14, defining factor II, "excuse the offender," it becomes possible to see that item 13 specifically refers to "rape" whereas the other variables defining this factor refer to forced sex/taking sex. Gender differences on this item are then at least intuitively understandable. These tentative suggestions are offered just to show the kind of inquiry that opens up when differentially functioning items are detected and studied.

BINARY ITEMS

The task in this section is to carry over the diagnostic devices in the previous section to binary items. The treatment here is limited to normal-ogive models without a pseudo-guessing parameter. Because a pseudo-guessing parameter is unaltered by scale transformations, the treatment applies equally to cases including such parameters. The normal-ogive is here preferred to a logistic model, for reasons that we demonstrate.

In Chapters 12 and 14 we found a direct connection between an item response model for binary data and the linear common factor model for quantitative item scores through item factor analysis. We assumed that a set of "underlying" quantitative response "tendencies" $X_1^*, \ldots, X_m^*$ follows the common factor model as in (15.1), and that the m binary variables $X_1, \ldots, X_m$ result from dichotomizations at threshold values $\tau_1, \ldots, \tau_m$ of the response tendencies. We then have, under normality assumptions,