Differential Testlet Functioning: Definitions and Detection

Author(s): Howard Wainer, Stephen G. Sireci and David Thissen

Source: *Journal of Educational Measurement*, Vol. 28, No. 3 (Autumn, 1991), pp. 197-219

Published by: National Council on Measurement in Education

Stable URL: http://www.jstor.org/stable/1434843

Accessed: 18/09/2013 22:10

Journal of Educational Measurement
Fall 1991, Vol. 28, No. 3, pp. 197–219

Differential Testlet Functioning: Definitions and Detection

Howard Wainer
Educational Testing Service
Stephen G. Sireci
Fordham University
David Thissen
University of North Carolina

It is sometimes sensible to think of the fundamental unit of test construction as being larger than an individual item. This unit, dubbed the testlet, *must pass muster in the same way that items do. One criterion of a good item is the absence of DIF—the item must function in the same way in all important subpopulations of examinees. In this article, we define what we mean by* testlet DIF *and provide a statistical methodology to detect it. This methodology parallels the IRT-based likelihood ratio procedures explored previously by Thissen, Steinberg, and Wainer (1988, in press). We illustrate this methodology with analyses of data from a testlet-based experimental version of the Scholastic Aptitude Test (SAT).*

It is often natural to think of the fungible unit of a test as a *testlet:* an interrelated and integrated group of items, always presented as a single unit (Wainer & Kiely, 1987). Historically, tests of skills such as reading comprehension have been constructed of testlets, or text passages followed by a number of interrelated questions (Thissen, Steinberg, & Mooney, 1989). But recent trends in test construction (Resnick, 1987; National Council of Teachers of Mathematics, 1989) emphasize a global view in the assessment of proficiency, and this trend toward focusing tests on a level larger than the item indicates a rich future for the use of testlets.

In parallel with this call for tests with greater construct validity has been a renewed emphasis on issues of test fairness. One aspect of fairness is the insistence that test items not function differentially for individuals of the same proficiency, regardless of their group membership. No DIFferential item functioning (DIF) is now a general desideratum. The area of study surrounding this desideratum has been defined formally and become referred to as DIF. A set of statistically rigorous and efficient procedures has been developed to

This work was supported by the Educational Testing Service through funding from the Program Research Planning Council and the New Possibilities Project. We would like to express our gratitude for this help. The work of the second author was done while he was an ETS Summer Predoctoral Fellow. The data reported on were made available to us by Nancy Feryok, John Fremer, and Ida Lawrence. We would like to express our gratitude to them as well as to Bill Angoff, Neil Dorans, Paul Holland, Charles Lewis, Robert Mislevy, Alicia Schmitt, William Stout, and Michael Zieky for their advice and comments on the work as it progressed.

detect and measure DIF. These generally fall into one of two classes; they are either based on latent variables (Thissen, Steinberg, & Wainer, 1988, in press) or on observed score (Holland & Thayer, 1988; Dorans & Holland, in press).

Procedures for DIF studies have traditionally focused on the item; indeed *item* is sometimes thought of as DIF's middle name. Yet, if future tests will be based on testlets, should we not generalize DIF procedures to suit this broader construct? The point of this article is to argue for precisely such a generalization and to provide two allied methods for accomplishing it. Because this article is statistical, we will not address the issues surrounding what one does with a testlet that is found to contain DIF. These issues are typically nonstatistical in nature involving decisions made on the basis of content and practicality. We leave such a discussion to other accounts.

Testlet DIF—An Inevitable Concept

The determination of DIF at the testlet level has three advantages over confining the investigation to the item. It allows:

(1) the analysis model to match the test construction,
(2) DIF cancellation through balancing,
(3) the uncovering of DIF that, because of its size, evades detection at the item level but can become visible with some aggregation.

Matching the Model to the Test

If a set of items were built to be administered as a unit, it is important that the items be analyzed that way. There are a variety of reasons for analyzing them as a unit, but underlying them all is the fact that, if one does not, one is likely to get the wrong answer. In the example described in a subsequent section, a four testlet test consisting of 45 separate items yields a reliability of .87 if calculated using traditional methods assuming 45 independent items. If one calculates reliability taking the within-testlet dependencies into account, the test's reliability is shown to be .76. These are quite different—note that Spearman-Brown (cited in Gulliksen, 1950/1987, p. 78) indicates that we would need to double the test length to yield such a gain in reliability (see Sireci, Thissen, & Wainer, 1991, for more details on this aspect). Other calculations (i.e., validity and information) are affected as well.

DIF Cancellation

Roznowski (1988), among others, has pointed out that because decisions are made at the scale or test level, DIF at the item level may have only limited importance. Therefore it is sensible to consider an aggregate measure of DIF. Small amounts of item DIF that cancel within the testlet would seem, under this argument, to yield a perfectly acceptable test construction unit. This is of critical importance in adaptive testing, less so with fixed format tests.

Humphreys (1962, 1970, 1981, 1986) has long argued that it is both inadvisable and difficult—very likely impossible—to try to construct a test of strictly unidimensional items. He suggests that to do so would be to construct a test that is sterile and too far abstracted from what would be commonly encountered to be worthwhile. He recommends the use of content rich (i.e., possibly multidimensional) items and suggests that, because multidimensionality is what causes DIF, we should control it by balancing across items. We agree with this. But balancing is not a trivial task. Surely such balancing needs to be done within content area and across the entire test. For example, it would be unfortunate if the items that favored one group were all at the end of the test. The concept of a testlet suggests itself naturally. Build the test out of testlets and ensure that there is no DIF at the testlet level. Lewis and Sheehan (1990) have shown that building a mastery test of parallel-form testlets provides a graceful solution to a set of thorny problems.

Cancellation of DIF could be accomplished in an adaptive testing situation without using testlets. However, it would involve accumulating DIF statistics of the items as they are to be administered and ensuring that the accumulation was zero when the test halted. This is almost surely possible without testlets, but it would certainly add a further burden to the item selection algorithm and item pool. Providing DIF-balanced testlets as the unit of test construction seems a much simpler strategy.

A final argument in support of examining DIF at the testlet level derives from the consideration of testlets that cannot easily be decomposed into items. For example, consider a multistep mathematics problem in which students get credit for each part successfully completed. Does it make sense to say that parts of such a testlet contain "positive subtraction DIF" and then "negative multiplication DIF?" Of course not. Instead, we must concentrate on the DIF of the problem as a whole. In some sense we do this now when we test an item's DIF. We do not record intermediate results and so do not know to what extent there is DIF on the component tasks required to complete the item. All we concern ourselves with is the final result.

It should be emphasized that by *cancel out* we mean something quite specific. We mean that there will be no DIF at every score level within the testlet. Exactly how we operationalize this goal, and what it means will be explicated and illustrated in the next sections.

Increased Sensitivity of Detection

It is possible (and, as we will demonstrate, even likely) to construct a testlet of items with no detectable item DIF. Yet the testlet in the aggregate does have DIF. The increased statistical power of dealing with DIF at the testlet level provides us with another tool to ensure fairness. This will be especially useful for those focal groups that are relatively rare in the examinee population and so are not likely to provide large samples during item pretesting.

Testlet DIF Detection—One Model, Two Methods

The polytomous IRT model we used was developed by Bock (1972). The basic notion is to fit the model to the data assuming that all testlets have the same parameters (no DIF) in the two populations of interest (*Reference* and *Focal*). We then fit the same model to the data allowing one testlet to have different parameters in each population (DIF) and compared the likelihood under each of the two situations. If the more general model did not yield a significant increase in the quality of the fit, we concluded that the extra generality was not needed and that the testlet in question had no DIF. This procedure was applied in the study of DIF by Thissen et al. (1988) using a more traditional dichotomous IRT model. Thissen et al. (1989) used Bock's polytomous model to fit testlets. Our testlet approach to DIF was almost exactly the one reported by Thissen et al., (in press) when we used the multiple choice model (Thissen & Steinberg, 1984) to examine differential alternative functioning (DAF). The step from DAF to testlet DIF was a small one.

Bock's 1972 Model

Suppose we have J testlets, indexed by j, where $j = 1, 2, \ldots, J$. On each testlet, there are m_j questions, so that for the jth testlet there is the possibility for the polytomous response, $x_j = 0, 1, 2, \ldots, m_j$. The statistical testlet scoring model posits a single underlying (and unobserved) dimension that we call latent proficiency and denote θ. The model then represents the probability of obtaining any particular score as a function of proficiency. For each testlet, there is a set of functions, one for each response category. These functions are sometimes called *item characteristic curves* (Lord & Novick, 1968), *item operating curves* (Samejima, 1969), or *trace lines* (Thissen et al., 1989). We shall follow Thissen et al.'s notation and nomenclature.

The trace line for score $x = 0, 1, \ldots, m_j$ for testlet j is

$$T_{\mathrm{jx}}(\theta) = \frac{\exp\,[a_{\mathrm{jx}}\theta + c_{\mathrm{jx}}]}{\sum_{\mathrm{k}=0}^{m} \exp\,[a_{\mathrm{jk}}\theta + c_{\mathrm{jk}}]}, \tag{1}$$

where the $\{a_k, c_k\}_j$, $k = 0, 1, \ldots, m_j$ are the item category parameters that characterize the shape of the individual response trace lines. The a_k are analogous to discriminations; the c_k are analogous to intercepts. The model is not fully identified, and thus we need to impose some additional constraints. It is convenient to insist that the sum of each of the sets of parameters equals zero—that is,

$$\sum_{k=0}^{m_j} a_{jk} = \sum_{k=0}^{m_j} c_{jk} = 0.$$

In this context, we reparameterize the model using centered polynomials of the associated scores to represent the category-to-category change in the a_k and the c_k:

$$a_{jk} = \sum_{p=1}^{P} \alpha_{jp} \left(k - \frac{m_j}{2}\right)^p \tag{2}$$

and

$$c_{jk} = \sum_{p=1}^{P} \gamma_{jp} \left(k - \frac{m_j}{2}\right)^p, \tag{3}$$

where the parameters $\{\alpha_p, \gamma_p\}_j$, $p = 1, 2, \ldots, P$ for $P \leq m_j$ are the free parameters to be estimated from the data. The polynomial representation has, in the past, saved degrees of freedom with no significant loss of accuracy. It also provides a check on the fit of the model when the categories are ordered. Although this model was developed for the nominal case, it can be used for ordered categories. If the categories are ordered, the a's must be monotonically ordered. (See the Appendix for proof.) As we show in the next section, the polynomial representation in this application saves degrees of freedom and indicates that the model provides a good representation of the data.

This version of Bock's model uses raw score within testlet as the carrier of information. It is possible that more information would be obtained by taking into account the pattern of responses within each testlet, but we felt that this simplification is appropriate for an initial foray into testlet DIF. Moreover, basing a test scoring algorithm on number right seems amply supported by general practice, especially as a first step.

In previous work, this model was fitted to a 4-passage, 22-item test of reading comprehension by Thissen et al. (1989), with m_j = (7, 4, 3, 8). The analysis followed an item factor analysis (Bock, Gibbons, & Muraki, 1988) that showed that a multifactor structure existed. The (at least) 4-factor structure found among these 22 items made the unidimensional assumption (conditional independence) of traditional IRT models untenable. After considering the test as four testlets and fitting Bock's nominal response model to the data generated by the almost 4,000 examinees, Thissen et al. compared the results obtained with what would have been the case if they had ignored the lack of conditional independence and merely fit a standard IRT model. They found two things: First, there seemed to be a slightly greater validity of the testlet derived scores when correlated with an external criterion. Second, the test information function yielded by the traditional analysis was much too high. This was caused by this model's not being able to deal with the excess intrapassage correlations among the items (excess after conditioning on θ). The testlet approach thus provided a more accurate estimate of the accuracy of the assessment. Through an obvious generalization, this same approach can be used to study testlet DIF.

Method 1: Internal Criterion

The basic data matrix of score patterns is shown in Table 1. In this example, there are four testlets with 10 possible score levels each [m_j = (10, 10, 10, 10)]; there are a maximum of 10^4 rows. In practice, there will be far fewer rows because many possible response patterns will not appear. The analysis follows what is done in item DIF situations: fitting one model to allow different values for the parameters of the studied testlet for the two groups and then comparing the -2loglikelihoods of that model with others that restrict the two groups' estimates in a variety of ways. Stratification/conditioning is done on θ, estimated for both groups simultaneously.

This method uses the test itself, including the studied testlet, to calculate the

TABLE 1
Arrangement of the Data for the Internal Analyses

Testlet Score Pattern				Total	Frequencies	
I	II	III	IV	Score	Reference	Focal
0	0	0	0	0	f_{R1}	f_{F1}
0	0	0	1	1	f_{R2}	f_{F2}
0	0	0	2	2	f_{R3}	f_{F3}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
s_I	s_{II}	s_{III}	s_{IV}	$\sum s_j$	f_{Ri}	f_{Fi}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	9	9	9	36	f_{RN}	f_{FN}

matching criterion. The question about whether or not to include the studied item has been carefully explored by Holland and Thayer (1988) who showed, for the Rasch model (the binary analog of this model), that not including the studied item in the criterion yields statistical bias under the null hypothesis. This was explored further by Zwick (1990) who confirmed this result for the Rasch model but not generally for other IRT models.

Using this method requires first fitting a completely unrestricted model—estimating all of the a_k and c_k separately for both the reference and the focal groups. Next, restricted versions of this model are estimated by approximating the values of the parameters as polynomial functions of score category (Equations 2 and 3). When an acceptably fitting parsimonious model is derived, we note the value of -2loglikelihood (asymptotically χ^2) for that model and then sequentially restrict the parameters for one testlet at a time to be equal across the two groups. We subtract the -2loglikelihood from the restricted model from the unrestricted and, remembering that the difference between two χ^2 statistics is also χ^2, we test that difference for significance; the number of degrees of freedom of the statistical test is equal to the number of parameters restricted. If it is not significant, we conclude that the extra flexibility gained by allowing different parameters for the focal and reference groups is not required—there is no DIF. If it is significant, we can further isolate where the DIF is located.

Eventually, one arrives at a determination of the most parsimonious representation. Interpreting the character of this representation allows us to detect testlet DIF. This is computationally expensive, with the cost of each run essentially linear in the number of response patterns observed. Of course, this cost is small relative to the cost of not detecting testlet DIF when it is there. The cost can be controlled substantially by reducing the number of possible response patterns. One way to do this is explored in the next section.

Method 2: External Criterion

The basic data matrix of score patterns is shown in Table 2. There is a matrix like this for both the Reference group ($G = R$) and the Focal group ($G = F$). For convenience, this example uses a six-item anchor yielding 2^6, or $N = 64$, possible matching levels.

This method uses an external criterion as the matching variable. This has been recommended as the practice of choice when a suitable external measure is available (Angoff, 1982, pp. 112–113; Thissen et al., in press). It cleanly avoids the issues surrounding what to do with the studied item when the matching criterion is internal as well as arguments of circularity. Of practical importance, the analysis focuses on a matrix 640 × 2: only 1,280 cells. This allows many items to be examined at only a modest cost in computer time. Contrast this with the parallel task utilizing an internal anchor that has 2-by-10^4, or 20,000 cells. The former analyses can be easily accommodated on a microcomputer; the latter is more comfortable on something larger, faster, and more expensive.

The strategy for accomplishing this analysis is quite similar to that described in the previous section. But there is one important extra step—the choice of the criterion items. We will not deal with the substantive aspects of that choice in this section; instead, we will focus on the psychometric characteristics used in the choice. The criterion items should: (a) be strongly related to the same underlying characteristic that is being measured by the testlets, (b) have steep slopes, (c) have their difficulties span the range of proficiency of the individuals taking the test, and (d) have no DIF. How many items are required? We have been successful with as few as three, but a more conservative stance (yielding protection against one of these items behaving poorly) would use five, six, or seven. We chose six in the example reported here—it worked very well indeed.

After choosing these special items (Thissen et al., in press, called these the *designated anchor*), each testlet takes its turn as the studied testlet. A saturated model is fitted, followed by suitably restricted ones. When the likelihood ratio

TABLE 2
Arrangement of the Data for the External Analyses

Criterion items						Testlet score									
1	2	3	4	5	6	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	f_{G10}	f_{G11}	f_{G12}	f_{G13}	f_{G14}	f_{G15}	f_{G16}.	.	.	.
0	0	0	0	0	1	f_{G20}	f_{G21}	f_{G22}	f_{G23}	f_{G24}	f_{G25}	f_{G26}.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
x_1	x_2	x_3	x_4	x_5	x_6	f_{Gi0}	f_{Gi1}	f_{Gi2}	f_{Gi3}	f_{Gi4}	f_{Gi5}	f_{Gi6}.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	1	1	1	1	1	f_{GN0}	f_{GN1}	f_{GN2}	f_{GN3}	f_{GN4}	f_{GN5}	f_{GN6}.			..

indicates that restricting the testlet's parameters to be equal across the two groups does not worsen the fit, we conclude that there is no DIF and move on to the next testlet. If it is significant, we continue our explorations to try to isolate the specific parameters that characterize the DIF.

Our experience with this methodology indicates that we obtain essentially the same results as with the more costly internal method. In the example described here, the computing time was about one third that used with the internal method. This is as expected because the size of the matrices used with the external criterion are about one third that using the internal criterion. With a larger sample of individuals, and hence more different response patterns, the difference would be more dramatic still.

Testlet DIF Applied—The NPP-V

The data analyzed here were part of the Spring, 1989, field testing of the New Possibilities Prototype test (NPP), an experimental version of the Scholastic Aptitude Test (SAT). This field testing represents an ongoing collaborative effort by the College Board and the Educational Testing Service (ETS) that is designed to investigate possible enhancements to the current SAT. The verbal section of the NPP, the NPP-V, includes longer reading passages than the SAT and has more items associated with each passage. The form of the NPP-V analyzed here consists of 75 multiple-choice items, 45 of which correspond to four long reading passages. These reading passages have 12, 13, 10, and 10 corresponding items respectively. We shall henceforth refer to these four passages as Testlets I, II, III, and IV. A more complete description of the NPP is not currently available, but one will be within the year. The analyses were based on 4,028 high school students: 2,216 females and 1,812 males.

Analysis Preliminaries

All analyses were done using *MULTILOG Version 6.0* (Thissen, 1991); it allows the mixing of item types within the same analysis that is crucial for the use of an external anchor of dichotomous items. It also allows the imposition of equality constraints that is necessary to obtain the likelihood of restricted (no DIF) models.

The maximum number of categories that the current version of *MULTILOG* allows for any polytomous model is 10. This limit required that we collapse some of the response categories in Testlets I and II. Because categories with very few entries provide poor parameter estimates, we found that little power was lost, and indeed some stability was gained[1] by combining some extreme score categories. Testlet I's 12 categories were reduced to 10 by combining Score Groups 0 and 1 into a new group labeled 0 and Categories 11 and 12 into a single category labeled 9. Testlet II's 13 score categories were similarly reduced by combining the three lowest (0, 1, 2) into Category 0 and the two highest into Category 9.

Previous experience (Thissen et al., 1989) has shown that trace lines for essentially chance scores are sufficiently similar to one another so that they can be combined with no loss of information. These are all five choice items, and we would expect chance performance on such testlets to yield scores around 2. The number of individuals in the highest categories was sufficiently small so that the judicious melding of those score categories would yield nothing but statistical stability. Thus we felt that this accommodation to the limits of the current version of *MULTILOG* would not influence our results.

Results of Method 1: Internal Anchor

The analysis began by fitting a completely unconstrained model to the data. This allowed each testlet to be fitted separately by sex. The polynomials described in Equations 2 and 3 were of ninth degree. We subsequently found that for the four testlets fitted we never needed greater than third degree polynomials and that often linear or quadratic functions gave wonderful fits.

Shown in Figure 1 are the fitted (line) and actual (points) values for the a_k values for Testlet I for males. We reproduce these here to show the closeness that (in this case) a quadratic approximation has to the actual data (this is the worst fitting set of parameters in this study). Shown in Figure 2 are the fitted and actual values for the c_k for Testlet I obtained from the male examinees. Once again this is the worst fitting polynomial that we found. Moreover, the constrained values are depicted with the fitted line, the unconstrained by the plotted points.

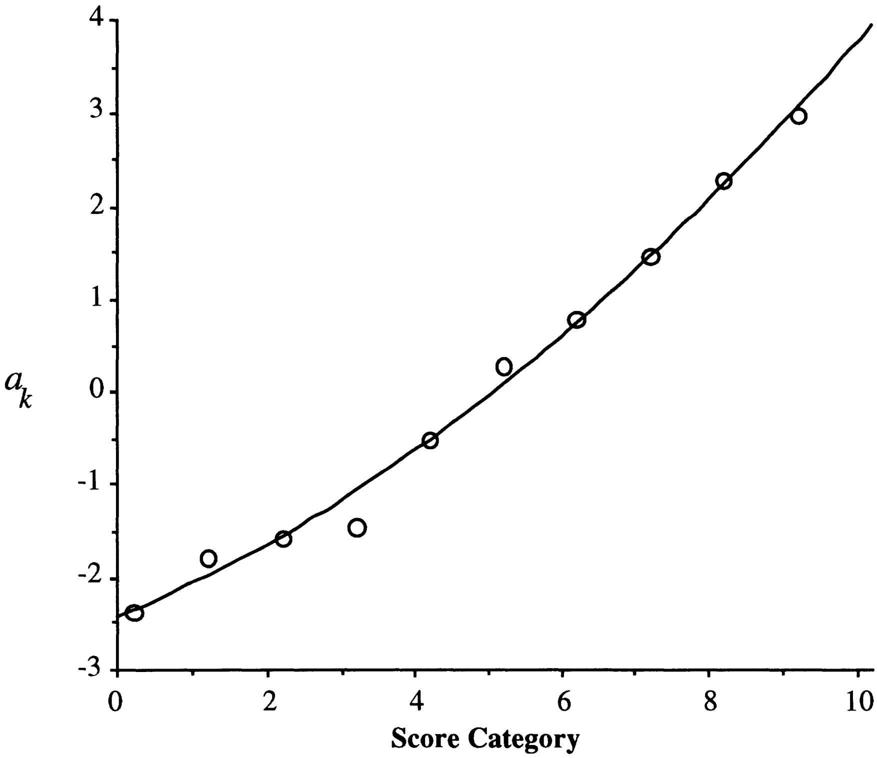


FIGURE 1. *The values of a_k for Testlet I for males, plotted against score category*

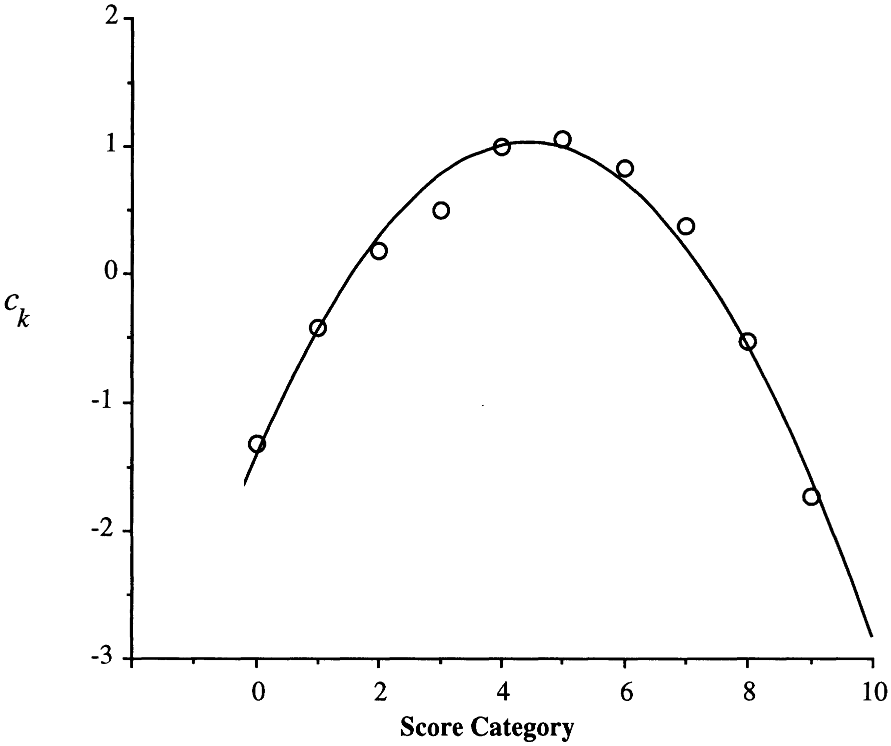


FIGURE 2. *The values of c_k for Testlet I for males, plotted against score category*

After determining the proper level of generality for the polynomial representation for the parameters of each of the testlets (examined separately for men and women), we began fitting a sequence of hierarchically nested models. We started with a completely unconstrained model (commonly termed *fully saturated*) in which each testlet had different parameters for males and females. Next we fit a completely constrained model in which the parameters for each testlet administered to males were constrained to be equal to the corresponding parameters of that testlet administered to females. We then moved from the constrained to the unconstrained model in directed steps. The results are shown in Table 3 and summarized graphically in Figure 3.

We can quickly see that the No DIF model can be rejected out of hand. The next sequence of four models tests whether the DIF can be isolated within a single testlet. The answer is no, but we get some useful information about what is going on. There are major decreases in misfit when Testlets I and II are allowed to show DIF, but allowing Testlets III and IV to have separately estimated parameters by sex yields no increase in the quality of the fit. It appears that it is likely that the DIF is located in Testlets I and II. The next row of Table 3 shows that, when we fit a model that restricts Testlets III and IV to be equal in both groups but allows separate estimation in Testlets I and II, the fit is not significantly different than the unconstrained model. Plotting the

TABLE 3
Summary of Search for Testlet DIF with an Internal Anchor

Model	–2Loglikelihood	# of Free Parameters	Difference $\chi 2$	*df*	*P*
Unconstrained	8412	35			
No DIF	8620	18	208	17	<.001
Just I DIF	8511	22	99	13	<.001
Just II DIF	8466	23	54	12	<.001
Just III DIF	8616	21	204	14	<.001
Just IV DIF	8617	23	205	12	<.001
I & II DIF	8421	27	9	8	0.4
I (C) & II DIF	8425	25	13	10	0.2

parameters separately estimated for Testlets I and II suggested that both the discriminations and thresholds for Testlet II were quite different for the two sexes, but for Testlet I only the c_k seemed to be different. Figures 4 and 5 show these plots for Testlet I.

The information in Figure 4 led us to constrain the a_k in Testlet I. Thus we arrived at the final model that indicated DIF in Testlet I only in location parameters, in Testlet II in both discrimination and location parameters, and

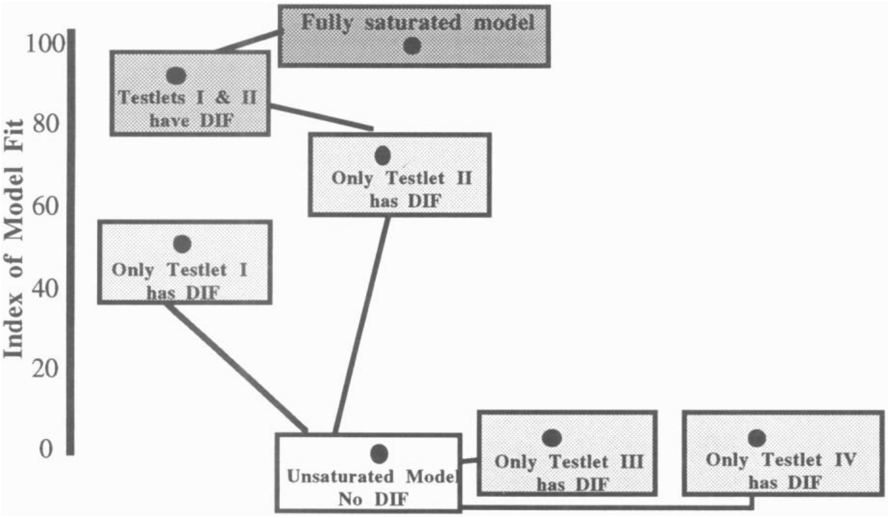


FIGURE 3. *Summary of the search for testlet DIF using an internal anchor*

Note. Each model is plotted at its level on an index of model fit (after Bentler & Bonett, 1980), ranging from 0 for the model with no DIF to 100 for the model with DIF for all testlets.

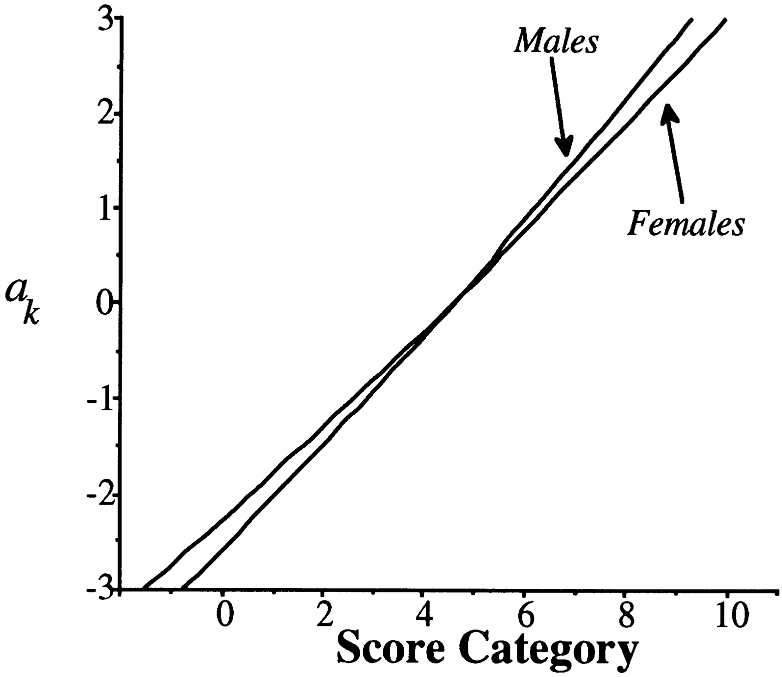


FIGURE 4. *The values of a_k for Testlet I for males and females compared, plotted against score category*

no DIF in either Testlet III or IV. Now that we had located the DIF, it remained that we try to understand it. Plots of parameters of a polytomous IRT model are not always easy to figure out. The next step is to examine the trace lines associated with these parameters.

In the interests of parsimonious presentation, we will not reproduce here the trace lines for all of the testlets; instead, we will focus on Testlet I. We do this to illustrate a variety of points; key among these is the size of DIF detectable with this methodology and this sample size. In the upper and lower panels of Figure 6 are the trace lines for Testlet I for males and females respectively. They look remarkably similar; however, the trace lines for the males are shifted to the left, relative to the female trace lines, for the higher testlet scores. The location of the shift in the trace lines shows where there is DIF, and the amount of shift indicates the amount of DIF. Evaluating the size of the DIF requires weighting the differences by the proficiency distribution of the focal group (Wainer, in press). The DIF is difficult to see in Figure 6; fortunately, there is another way to examine the result.

Each of the trace lines in Figure 6 indicates the conditional probability of an individual's being in that score group: $[P(x = k|\theta)$ for $k = 0, 1, \ldots, 9]$. Plotting the expected conditional score group, $[E(x|\theta) = \Sigma xP(x\ |\theta)]$, reduces the sheaf of 10 curves for each sex to a single function. The expected score group is very close in both form and spirit to Lord's (1980) recommendation regarding the use of expected true score. Shown in Figure 7 is a plot of the expected score

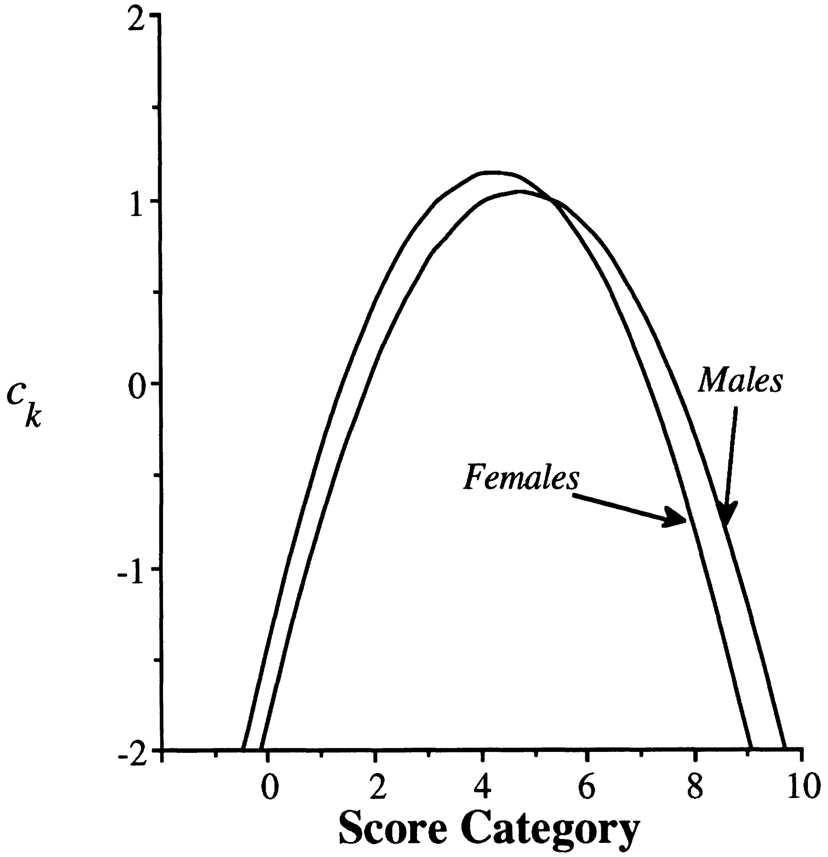


FIGURE 5. *The values of c_k for Testlet I for males and females compared, plotted against score category*

groups for males and females. The direction of the advantage is clear. If we subtract the females' curve from the males' curve, we obtain a clear depiction of the size of the DIF (see Figure 8). From this plot, we see that the maximum advantage is about a half point on a 10 point scale (about 5%).

Before concluding this section, let us examine the size and direction of the DIF found in Testlet II using plots of expected score category. The difference plot for Testlet II is shown in Figure 9. We see immediately that this time the DIF is in favor of females but that the advantage disappears at higher proficiency levels. The effect of unequal discriminations (a_k) is apparent. We also note (in Figure 9) that the maximal advantage to females is about one point (out of 10), and it is centered at about the center of the proficiency distribution. This is roughly twice the DIF seen in Testlet I.

The size and direction of the testlet DIF detected provides a sense of the statistical power of this methodology. Testlets III and IV had no detectable DIF. By that we mean that any DIF that might have existed within those two testlets was smaller than that shown here. It should be emphasized that there

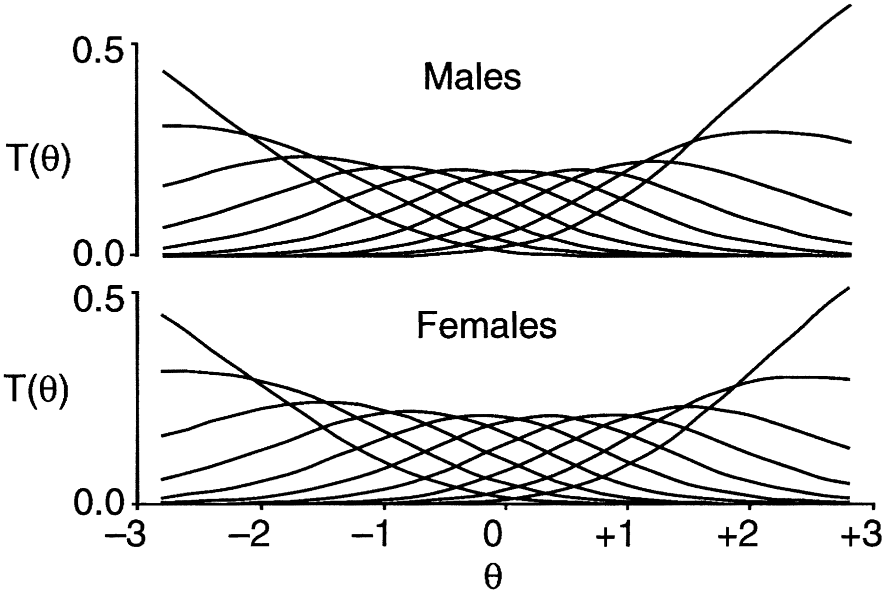


FIGURE 6. *Trace lines for the 10 response categories for Testlet I*

Note. The modes of the trace lines are in the order of the score-group categories 0–9.

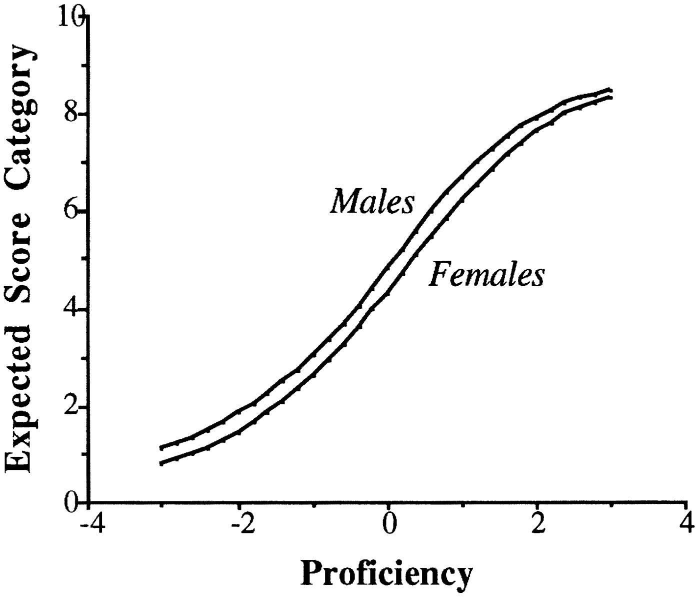


FIGURE 7. *Expected score category on Testlet I plotted against proficiency for males and females*

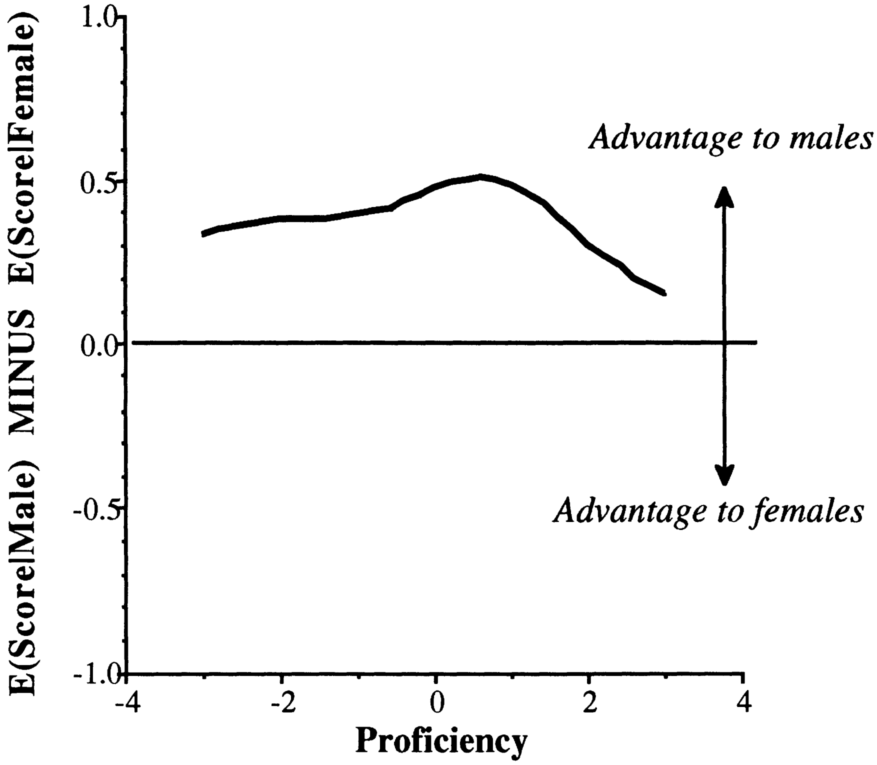


FIGURE 8. *The difference between males and females in expected score category on Testlet I plotted against proficiency*

were only a few items in any of the testlets that showed significant DIF when screened individually.[2] Thus, examining entire testlets for DIF as a whole has provided us with a tool with increased sensitivity. However, the cost of this increased statistical power has been a substantial increase in the conceptual and computational complexity. In the next section, we show how a much simpler methodology gets us essentially identical results.

Results of Method 2: External Anchor

When we use an internal anchor, the basic data matrix is potentially very large indeed; $10^4 \times 2$ for four testlets with 10 score categories. The size of the analysis problem goes up exponentially with the number of testlets—10^5 for five testlets, 10^6 for six, etc. This can be controlled and sharply reduced through the use of a fixed external anchor. In this case, we chose six dichotomous items from among the 30 multiple-choice items that were also on the NPP-V. We chose these items very carefully. They were the items with the lowest DIF (measured using the Mantel-Haenszel statistic) that spanned the range of the proficiency distribution. We also tried to choose items that had good discrimination.

These six anchor items were fitted with the three-parameter logistic IRT

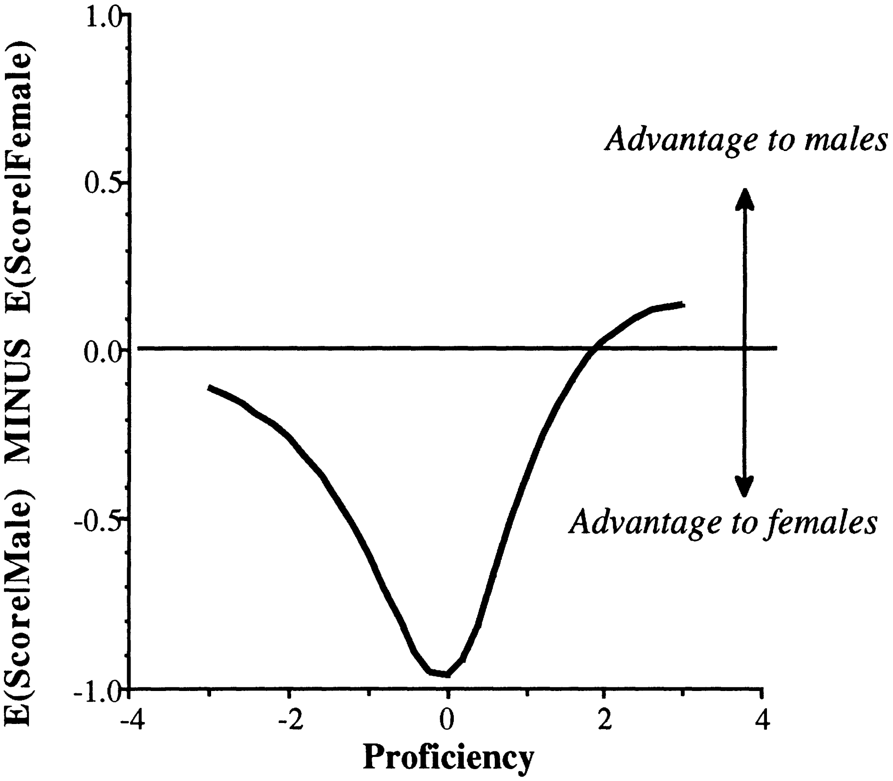
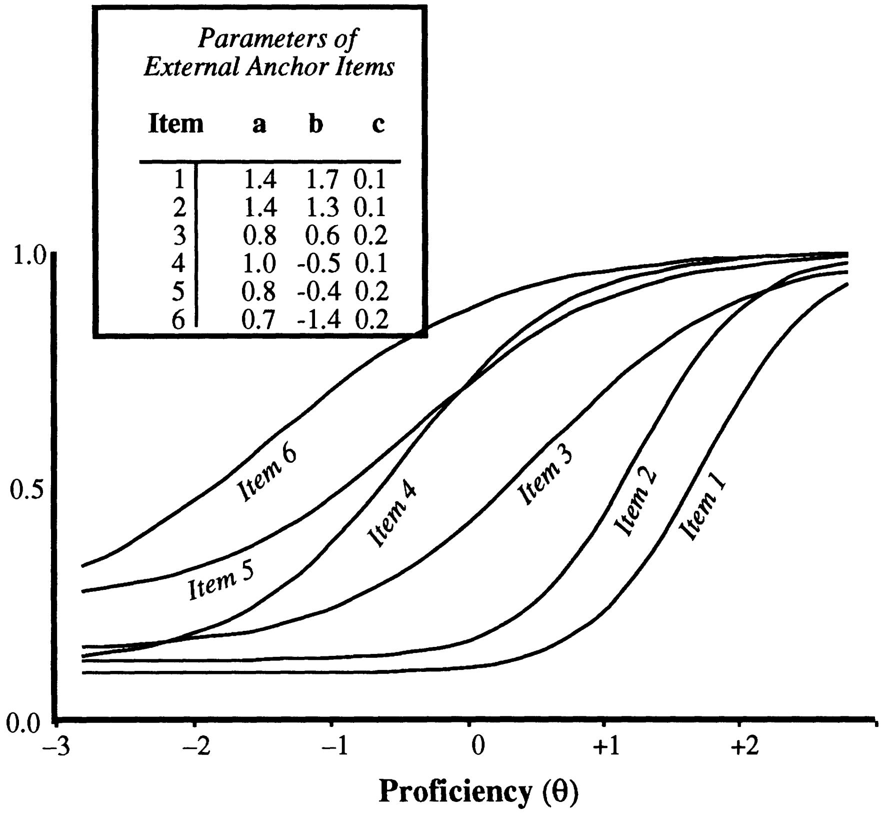


FIGURE 9. *The difference between males and females in expected score category on Testlet II plotted against proficiency*

model (3-PL) in the course of the DIF analyses. In Figure 10 are the estimated parameters for these items and the plots of their trace lines.

Once the anchor items were chosen, we followed the procedure described earlier. This required appending each testlet in turn as the studied testlet to the six-item anchor. Then we fit an unconstrained model allowing the testlet to have different parameters for the focal and reference group as well as a model in which the testlet's parameters were constrained to be equal in the two groups. Once again we looked at the likelihood ratio, and if there was no significant increase in fit with the relaxation of the equality constraints we concluded that there was no DIF. The results of these analyses are summarized in Table 4.

It is clear that the conclusions that can be drawn from the results shown in Table 4 are similar to those drawn from the internal analyses; this adds empirical support to the practical reasons for using a short external anchor of multiple-choice items to stratify the examinee population. Testlet I shows DIF only in the threshold parameters (the c_k)—note that the likelihood ratio χ^2 comparing a model restricting just a_k to a model with no restrictions is 2 on 2 degrees of freedom. Testlet II shows DIF; Testlets III and IV show no DIF.



Parameters of External Anchor Items

Item	a	b	c
1	1.4	1.7	0.1
2	1.4	1.3	0.1
3	0.8	0.6	0.2
4	1.0	-0.5	0.1
5	0.8	-0.4	0.2
6	0.7	-1.4	0.2

FIGURE 10. *Trace lines and parameters for the six external anchor items*

Plots of the trace lines for the testlets estimated within the context of an external anchor are virtually identical to those obtained with the far more computationally intensive internal anchoring procedures. Each estimation run here requires analysis of a 64 × 10 × 2 table (1,280 cells). Two runs are required for each testlet. The internal anchor is far more complex. As the number of testlets increases, computing time using an external anchor increases linearly, whereas using an internal anchor increases time exponentially. Our experience so far suggests that if a good external anchor can be constructed one would be foolish not to use it.

Restrictions surrounding security of operational test forms preclude us from any extensive discussion of the content of the testlets analyzed in this article. However, the passage associated with Testlet II (DIF favoring women) was an extended description of a visit to a grandmother. Testlet IV (no DIF) involved excerpts from speeches by Pericles and Abraham Lincoln.

Testlet DIF Whither—A Discussion of What's Next

It is to the benefit of large testing organizations to look for DIF and not find any. In statistical terminology, not finding DIF means being unable to reject the null hypothesis. That is, we assume that there is no DIF and after considering

TABLE 4
Summary of Search for Testlet DIF with an External Anchor

Model	–2Loglikelihood	# of Free Parameters	Difference $\chi 2$	*df*	*P*
Testlet I					
No DIF	1256	23			
DIF in *c*'s only	1236	25	20	2	<.0001
DIF in *c*'s and *a*'s	1234	27	22	4	<.0001
Testlet II					
No DIF	1396	24			
DIF	1312	29	84	5	<.0001
Testlet III					
No DIF	1206	22			
DIF	1204	25	2	3	0.68
Testlet IV					
No DIF	1239	24			
DIF	1231	29	8	5	0.12

the evidence decide that we cannot reject that hypothesis. It is easy to accept the hypothesis that there is no DIF. To accomplish this, one merely has to run poor studies with smallish sample sizes and use weak statistical models. Thus, to be credible, a finding of no DIF must be accompanied by a careful study with as large a sample size as can be found. The study must also use the strongest (i.e., the most efficient) statistical model available to analyze these data.

The history of DIF procedures, described by Angoff (in press), illustrates how statistical methods were initially developed to match heuristic ideas about what ought to be measured. This was, properly, the most important initial concern, with niceties such as statistical power being left for later. In the past few years, two classes of powerful models for detecting and measuring DIF have become available. Dorans and Holland (in press) provide a thorough description of two quite similar procedures (one based on standardization and the other on the Mantel-Haenszel statistic). These methods are nonparametric in that they do not attempt to model response likelihoods. Both methods are statistically efficient and inexpensive to compute.

Thissen et al. (in press) describe methods that utilize a likelihood ratio of two models to detect DIF. Statistical theory predicts that these methods are asymptotically optimal when the IRT model that is assumed to underlie the individual item responses is appropriate. In this chapter, the authors demonstrate how the methodology generalizes easily to study patterns of differential response among the item's distractors. This generalization is achieved through the use of a polytomous IRT model and results in what the authors call a methodology for studying DAF. This powerful new tool is shown to be helpful in diagnosing the misfunctioning of an item after DIF has been detected.

In the current presentation, we have generalized DAF procedures to allow the detection of testlet DIF. We have shown that this generalization accomplishes a variety of worthy goals. We showed that:

1. It characterizes the statistical character of the test more accurately than is the case with any model that does not acknowledge the clustered structure of the test's items. We illustrated this when we pointed out that by not modeling the testlet structure the reliability of the test was overestimated by an amount equivalent to a test of doubled length.
2. Testlets made up of what appeared to be exemplary items (both Testlets I and II), exhibited significant sex DIF when the testlets were considered in toto. This increased statistical power is especially important when we study the suitability of newly developed items for subpopulations of examinees who show up only seldom in test samples.
3. Testlets constituted of items with modest DIF in both directions can still be fair at all score and proficiency levels (Testlet IV).

We believe that because the current weltanschauung points test development toward tests composed of larger tasks it is well that we have the statistical tools to properly deal with such tests. The concept of the testlet and the associated psychometrics is a big step in that direction.

We recognize that procedures based on the fitting of hierarchical IRT models and the examination of likelihood ratios do not meld well with the economic stringencies of mass testing. Imagine the resources required for a detailed examination of the thousands of items required for an adaptive item pool! It would surely be better if something more computationally parsimonious could be found. Paul Holland often promoted the Mantel-Haenszel procedure by exclaiming, "10¢ an item!" Perhaps for a 13-item testlet he would be content with achieving the goal of $1.30 a testlet. Using an internal anchor does not approach this goal, although it does allow a level of detail in the investigation that has not been approached yet with other methods. The external anchor methodology is much more practical and sacrifices little or none of the power of the internal method. It also illustrates the single greatest strength of IRT-based methods—it can stratify individuals on a short anchor (Bock, in press). Nonmodel-based methods like those utilizing the Mantel-Haenszel or standardization procedures stratify examinees on their raw score. This works fine when a test is long enough to do this reliably. But on short tests reliable stratification requires utilizing information from response patterns. Some IRT models do this and so yield the accuracy of result we illustrate here.

We believe that testlet-based generalizations of the Mantel-Haenszel procedure can be usefully applied. For example, one obvious generalization would stratify individuals by score and then within-score stratum would calculate the contingency table of testlet score-by-group membership. Current usage with dichotomous items yields a 2 × 2 matrix (correct-incorrect × Focal-Reference); this generalization would yield an $m_j \times 2$ matrix. The statistic would then (as now) be calculated under the hypothesis of no interaction. Of course, this sort of generalization could be used not only with dichotomous items but also with several groups—testing for DIF in all (say n) focal groups at once. Why hasn't it been used this way? The answer relates to the statistical fact that one-degree-of-freedom tests are the most powerful. Thus, we achieve a more sensitive detection instrument if we do a series of one-degree-of-freedom tests rather than a single $(n - 1)$ degree-of-freedom test. To match this attitude in using the Mantel-Haenszel procedure to detect testlet DIF, one might want to collapse score categories to just two (perhaps above and below average). Then one might grind on with the usual Mantel-Haenszel procedure. Contrast this more extreme approach with our practice of collapsing to 10 categories. We suspect that we could obtain somewhat more power—but at a cost of understanding exactly where the problem lies.

We considered these arguments in our development of the methodology presented here. We believe that we have arrived at a sensible compromise between power and delicacy. Anyone doubting this should consider the size of the differences that were uncovered as being statistically significant (see Figure 7). The samples we have used are realistic for most practical situations to reliably detect rather small amounts of DIF using an anchor of only six items. We believe that such an anchor represents a method of sufficient power for most applications.

Notes

[1] Coefficient α was higher for summed scores with the extreme score categories collapsed than it was in the original data.

[2] Actually there were a couple of items in Testlet IV that demonstrated a modest amount of DIF. However, these items were counterbalanced by others that showed small DIF in the other direction. As we have shown, this counterbalancing, whether intentional or not, was effective in yielding a testlet that has no significant DIF in any score category at any value of proficiency.

APPENDIX

On an item that is scored in an ordered scale such as 1, 2, . . . , m, we would like the odds of being in a higher score category to be greater for an examinee of greater proficiency than for one with less.

Stated symbolically,

$$\frac{P(x = j \mid \theta = \theta_1)}{P(x = k \mid \theta = \theta_1)} > \frac{P(x = j \mid \theta = \theta_2)}{P(x = k \mid \theta = \theta_2)}, \tag{A1}$$

where x is the observed score, $j > k$, θ is proficiency, and $\theta_1 > \theta_2$.

Using shorthand notation $P(x_j \mid \theta_1)$ to mean $P(x = j \mid \theta = \theta_1)$ and rearranging allows us to rewrite inequality (A1) as

$$\frac{P(x_j \mid \theta_1)P(x_k \mid \theta_2)}{P(x_k \mid \theta_1)P(x_j \mid \theta_2)} > 1. \tag{A2}$$

(continued on p. 217)

Taking logs yields

$$\ln [\mathrm{P}(x_j | \theta_1)] + \ln [\mathrm{P}(x_k | \theta_2)] - \ln [\mathrm{P}(x_j | \theta_2)] - \ln [\mathrm{P}(x_k | \theta_1)] > 0. \qquad \text{(A3)}$$

If we model the probabilities with Bock's (1972) formulation for a categorical model (Equations 1 and 2), we find that

$$\ln [\mathrm{P}(x_j | \theta_1)] = a_j \theta + c_j - \ln [\textit{denominator}].$$

After substituting this in inequality (3), we find that the denominators cancel out, and we are left with

$$(a_j \theta_1 + c_j) + (a_k \theta_2 + c_k) - (a_j \theta_2 + c_j) - (a_k \theta_1 + c_k) > 0. \qquad \text{(A4)}$$

Rearranging and canceling yields

$$a_j - a_k > 0$$

or

$$a_j > a_k \qquad \text{for} \qquad j > k. \qquad \text{(A5)}$$

This tells us that accomplishing our goal requires an ordering of the *a* parameters.

Thus, the practice of fitting a monotone function to the initially estimated *a*'s not only provides a more parsimonious model but ensures that Bock's nominal model yields satisfactory results for ordered categories of scoring.

Our thanks to Paul Holland and Charles Lewis who, respectively, pointed out the mathematical relationship described above as well as its importance.

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 item bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H. (in press). Perspectives on the theory and application of differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29–51.

Bock, R. D. (in press). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Dorans, N. J., & Holland, P. W. (in press). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gulliksen, H. O. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates. (Original work published in 1950)

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475–483.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23–32). Seattle: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum Press.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367–386.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Resnick, L. B. (1987). *Education and learning to think.* Committee on Mathematics, Science, and Technology Education, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Roznowski, M. (1988). Review of test validity. *Journal of Educational Measurement, 25,* 357–361.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 4,* (17, Pt. 2).

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Thissen, D. (1991). *MULTILOG user's guide (Version 6)* [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items, *Psychometrika, 49,* 501–519.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247–260.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (in press). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (in press). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–198.

Authors

HOWARD WAINER is Principal Research Scientist, Educational Testing Service, 21-T, Princeton, NJ 08541. *Degree:* PhD, Princeton University. *Specializations:* statistical graphics, psychometrics, and computerized testing.

STEPHEN G. SIRECI is Psychometrician, American Institute of Certified Public Accountants, Examinations Division, 1211 Avenue of the Americas, New York, NY 10036. *Degrees:* BA, MA, Loyola College; PhD Candidate, Fordham University. *Specializations:* psychometrics and classification.

DAVID THISSEN is Professor of Psychology, University of North Carolina, CB 3270, Davie Hall, Chapel Hill, NC 27599-3270. *Degrees:* BA, St. Louis University; PhD, University of Chicago. *Specializations:* quantitative psychology and measurement.