DOCUMENT RESUME

ED 468 956 TM 034 443

AUTHOR	Wainer, Howard
TITLE	A Testlet-Based Examination of the LSAT. Statistical Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-SR-93-03
PUB DATE	1994-03-00
NOTE	30p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Item Response Theory; Reading Comprehension; Reading Tests; *Reliability; Test Bias; *Test Construction
IDENTIFIERS	*Law School Admission Test; *Testlets

ABSTRACT

This study examined the Law School Admission Test (LSAT) through the use of testlet methods to model its inherent, locally dependent structure. Precision, measured by reliability, and fairness, measured by the comparability of performance across all identified subgroups of examinees, were the focus of the study. The polytomous item response theory model used was developed by R. Bock (1972), and for the detection of testlets, likelihood ratio tests were used. All the analyses were performed on all four sections of two parallel forms of the LSAT. The analyses show that the testlet structure of the Reading Comprehension and Analytical Reasoning sections of the LSAT has a significant effect on the statistical characteristics of the test. The testlet-based reliability of these two sections is considerably lower than that was previously calculated under the assumption of local conditional independence. The discovery suggests that, at a minimum, any reporting of section reliability ought to be modified to reflect current knowledge, and analyses of current and future forms of these sections of the LSAT ought to model the testlet structure explicitly before calculating section reliability. (Contains 2 figures, 6 tables, and 64 references.) (SLD)

LSAC RESEARCH REPORT SERIES

TM

A Testlet-Based Examination of the LSAT

Howard Wainer

Law School Admission Council
Statistical Report 93-03
March 1994

A Publication of Law School Admission Council/Law School Admission Services

LSAC RESEARCH REPORT SERIES

A Testlet-Based Examination of the LSAT

Howard Wainer

Law School Admission Council
Statistical Report 93-03
March 1994

A Publication of Law School Admission Council/Law School Admission Services

The Law School Admission Council is a nonprofit association of United States and Canadian law schools. Law School Admission Services administers the Council's programs and provides services to the legal education community.

This study is published and distributed by the Law School Admission Council (LSAC) and Law School Admission Services (LSAS). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of LSAC/LSAS.

Table of Contents

I. Introduction[1]

The LSAT is composed of four sections. There is one section each of Reading Comprehension and Analytical Reasoning questions and two sections of Logical Reasoning. Each Logical Reasoning section is made up of about 25 independent items[2] and as such their statistical characteristics can be evaluated with standard procedures (e.g., Gulliksen, 1950, 1987). Reading Comprehension and Analytical Reasoning are not made up of independent items. In both cases their items are clustered around common stems. In Reading Comprehension there are 28 items that address one of four reading passages. In Analytical Reasoning there were 24 items related to one of four situations. In no case were there fewer than 5 nor more than 8 items attached together through a common stem. The common stem that these items shared provides strong prima facie evidence for the invalidity of the assumption of local conditional independence required for traditional analyses.

We reach this conclusion from observing that while the assumption of local independence is an apt description for many tests comprised of small, discrete items, there are other tests on which the items do not appear to be locally independent. Yen (1992) cataloged a large number of possible causes of local dependence; for instance, the several questions following each of a few reading comprehension passages typically exhibit local dependence, or the items at the end of a long (or somewhat speeded) test may similarly covary more strongly than an IRT model would predict.

Items exhibiting local dependence are, to some extent, redundant. To the extent that the response to the second item in a pair depends on (and therefore, can be predicted from) the response to the first item, beyond prediction from the underlying proficiency being measured, the second item provides less information about proficiency than would a completely (locally) independent item. Nevertheless, there are many reasons to include locally dependent groups of items on tests. Passages followed by questions are often thought to be the most authentic of reading comprehension tests, although the cluster of questions after each passage often exhibits local dependence. For this and other such situations, Wainer and Kiely (1987) introduced the concept of the *testlet*—a scoring unit within a test that is smaller than the test, comprised of several items that are usually assumed to be locally dependent. By applying multiple-categorical IRT models to responses of conditionally independent testlets, instead of to responses to the locally dependent items within the testlets, we have found that the statistics of the IRT model

[1]The work reported here is part of a long-term and on-going collaboration with David Thissen of the University of North Carolina at Chapel Hill. Our work is so intermingled that it is impossible to know whose is whose. This work shares prose with much of our collaborative work and as such perhaps should have quotation marks here and there. To do so is at least difficult, and may be impossible. Instead let me state clearly that many of the ideas expressed here as well as some of the prose may have originated in Thissen's word processor. He is aware of, and has approved, whatever plagiarism may have occurred.

[2]This is almost accurate, but not quite. In one section there were 20 independent items and two pairs of items. In the other section there were 21 independent items and two pairs of items. For all practical purposes one can think of these sections as having 24 and 25 independent items respectively. Our analyses modeled the pairing, but the results are sufficiently similar to those obtained by ignoring it, that we felt that such fastidiousness was unnecessary for any practical purpose.

generally perform better (see Sireci, Thissen, & Wainer, 1991; Thissen, 1993; Thissen & Steinberg, 1988; Thissen, Steinberg, & Mooney, 1989; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991).

This project has had as its goal the careful examination of the LSAT through the use of these methods to model its inherent, locally dependent, structure. We focused our attention on two of the most important aspects of the test: its precision (as measured by its reliability), and its fairness (as measured by the comparability of its performance across all of the identified subgroups of examinees taking the exam). The explicit modeling of the testlet structure of the LSAT has no drawbacks, other than a slightly increased complexity of calculation, and important advantages. If we are incorrect in our modeling, that is if the testlets which we believe to be locally dependent turn out to be independent, our results should mirror those obtained through traditional methods. If our assumptions about the structure are correct, the statistical characterization that emerges is a more accurate one.

In Section II of this report we describe the historical background and methodology associated with how reliability is measured traditionally and under a testlet-based formulation. In Section III we describe the methodology surrounding testlet-based DIF analysis. Section IV provides the results of these analyses on the two forms of the LSAT examined and Section V contains conclusions, discussion, implications, and some suggestions for future work. The appendices detail the results of all analyses.

II. Reliability

The fact that the reliability of tests built from testlets will be over-estimated by item-based methods is not newly discovered. It has been well-known for at least 60 years. Warnings about inflated reliability estimates are commonly carried in many introductory measurement textbooks. For example,

> *One precaution to be observed in making such an odd-even split pertains to groups of items dealing with a single problem, such as questions referring to a particular mechanical diagram or to a given passage in a reading test. In this case, a whole group of items should be assigned intact to one or the other half. Were the items in such a group to be placed in different halves of the test, the similarity of the half scores would be spuriously inflated, since any single error in understanding of the problem might affect items in both halves.*—Anastasi (1961, p. 121; included in all subsequent editions as well, see 6th edition, p. 121)

> *In some cases, several items may be unduly closely related in content. Examples of this would be a group of reading comprehension items all based on the same passage... In that case, it will be preferable to put all items in a single group into a single half-score... this procedure may be expected to give a somewhat more conservative and a more appropriate estimate of reliability...*—Thorndike (1951; p. 585)

> *Interdependent items tend to reduce the reliability. Such items are passed or failed together and this has the equivalent result of reducing the length of the test.*—Guilford (1936; p. 417)

These warnings were generated by an earlier exchange about the potential usefulness of the Spearman-Brown prophecy formula. This exchange is summarized by Brown and Thomson (1925), who reported criticisms of the mathematician W. L. Crumm (1923). Crumm felt that the requirements of Spearman-Brown were too stringent to expect them to be met in practice. Holzinger (1923) reported some empirical evidence supporting the Spearman-Brown formula, and then Truman Kelley (1924) provided a fuller mathematical defense of Spearman-Brown. Of interest here is Kelley's (1924) comment, "If two

or more exercises contain common features, not found in the general field, then the Spearman-Brown r_{11} will tend on this account to be too large" (p. 195).

All of these are in the context of even-odd split half reliability; but coefficient α averages all of the split-half reliabilities. Thus if some of the split-half reliabilities are inflated by within-passage correlation, so too will be α.

The APA *Standards for Educational and Psychological Tests* (section D5.4) deemed treating testlets in a unitary matter "essential." It stated

> *If several questions within a test are experimentally linked so that the reaction to one question influences the reaction to another, the entire group of questions should be treated preferably as an "item" when the data arising from application of split-half or appropriate analysis-of-variance methods are reported in the test manual.*

For a variety of reasons and purposes, along with our colleagues, we (Thissen, Steinberg & Mooney, 1989; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991) have proposed (essentially) treating testlets as items; here we concentrate on the effects on the estimation of reliability that arise from failing to do so.

Methodology

Reliability is defined (Lord & Novick, 1968, p. 61) as the squared correlation ρ^2XT between observed score (X) and true score (T). This can be expressed as the ratio of true score variance to observed score variance, or after a little algebra

$$\rho^2XT = \frac{\sigma^2X - \sigma^2e}{\sigma^2X} \qquad (1)$$

where the subscript *e* denotes error. In this report, we use coefficient α (see Lord & Novick, p. 204) to estimate the reliability of the summed scores of the traditional test theory.

There are many ways to calculate reliability, but in traditional test theory reliability takes a single value that describes the average error variance for all scores. [Traub and Rowley (1991, p. 40) use the notation Ave (σ^2_e) in place of σ^2_e to emphasize that only an average estimate of the error variance is considered in true score reliability theory.] In contrast to this simplification, measurement precision in an IRT system can be characterized as a function of proficiency (θ); therefore precision need not be represented by a single overall "reliability." Precision in an IRT system is usually described in terms of $I(\theta)$, the information function, the conditional error variance σ^2_{e*}, or the standard error σ_{e*}; these also vary as functions of θ. Although measurement error may vary as a function of proficiency, Green, Bock, Humphreys, Linn and Reckase (1984) observe that it can also be averaged to give "marginal reliability" comparable to that of the traditional theory. The marginal measurement error variance, σ^{-2}_{e*}, for a population with proficiency density $g(\theta)$ is

$$\sigma^{-2}e* = \int \sigma^2 e * g(\theta)\ d\theta\ , \qquad (2)$$

where σ^{-2}_{e*} is the expected value of the error variance associated with the *expected a posteriori* estimate at θ, and the marginal reliability is

$$\bar{\rho} = \frac{\sigma^2\theta - \sigma^{-2}e}{\sigma^2\theta} \quad (3)$$

Here, the integration (or averaging) over possible values of θ takes the place of the traditional characterization of an average error variance. The value of σ^{-2}_{e*} is the average of the (possibly varying) values of the expected error variance, σ^2_{e*}; if many values of σ^{-2}_{e*} were tabulated in a row for all the different values of proficiency (θ), σ^{-2}_{e*} would be that row's marginal average. Therefore, the reliability derived from that marginal error variance is called the marginal reliability, and denoted $\bar{\rho}$ to indicate explicitly that it is an average. There is some loss of information in averaging unequal values of σ^2_{e*}. But such marginal reliabilities for IRT scores parallel the construction of internal consistency estimates of reliability for traditional test scores.[3] In the remainder of this report, we use $\bar{\rho}$ to describe the reliability of estimates of latent proficiency (θ) based on IRT.

The surface analogy between equations 1 and 3 is slightly deceiving. The term σ^2_{e*} is the expected value of the error variance of the *expected a posteriori* estimate of θ as a function of θ, computed from the information function for the *expected a posteriori* estimate of θ—as Birnbaum (1968, pp. 472ff) illustrates, any method of test-scoring has an information function; here we use the information function for the *expected a posteriori* estimate. That is analogous to the error of estimation, σ^2_{ε}, in the traditional theory, not σ^2_{e*} (see Lord & Novick, p. 67). And σ^2_{θ} is the true population value of the variance of θ.

The definition of traditional reliability in terms of true score variance σ^2_T and the error of estimation σ^2_{ε} may be straightforwardly derived from Lord & Novick's (1968, p. 67) equation 3··8.4:

$$\theta_{\varepsilon} = \theta_T\sqrt{1-\rho^2XT} \ ,$$

which gives

$$\rho^2XT = \frac{\sigma^2T - \sigma^2\varepsilon}{\sigma^2T}$$

that is the classical analog of equation 3.

---

[3]Technically, Mislevy (personal communication, July 1990) has pointed out that if the information function is steeper than the proficiency distribution the integral in (2) can become unbounded. This is typically not a problem in applications where one approximates the integral with a sum. Yet one could imagine a case in which changing the bounds of the integral could drastically alter results. Mislevy's alternative is to integrate the information function directly and then invert. Specifically, substitute for (2) the expression

$$\bar{I}(\theta) = \int_{-\infty}^{\infty} I(\theta)g(\theta) = 1/\overline{\sigma^2e*}. \quad (2^*)$$

Mislevy's formulation has obvious theoretical advantages, but we have too little experience as yet to indicate under what circumstances the extra protection is needed.

An IRT Model for Testlets

While it is possible to approach testlet-analysis (the item analysis of testlets) using the tools of the traditional test theory, we have found the tools of IRT to be more useful. In the computation of the IRT index of reliability, $\bar{\rho}$, we follow Thissen, Steinberg, and Mooney (1989) in their use of Bock's (1972) model, in which we have J testlets, indexed by j, where $j = 1, 2, ..., J$. On each testlet there are m_j questions, so that for the jth testlet there is the possibility for the polytomous response $x_j = 0, 1, 2, ..., m_j$. The statistical testlet scoring model posits a single underlying (and unobserved) dimension which 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 response functions (IRFs) (Holland, 1990).

The IRF for score $x = 0, 1, ..., m_j$, for testlet j is

$$\frac{\exp[\alpha_{jx}\theta + c_{jk}]}{\sum_{k=0}^{m} \exp[\alpha_{jx}\theta + c_{jk}]} \qquad (4)$$

where $\{a_{jx}, c_{jk}\} j,k = 0,1,...,mj$ are the item category parameters that characterize the shape of the individual response functions. If this model yields a satisfactory fit, information is calculated and inverted to yield estimates of the error variance function. Error variance is relative to the variance of θ, whose distribution $g(\theta)$ is fixed as $N(0,1)$ to help identify the model. The integration indicated in (2) can then be carried out and $\bar{\rho}$ calculated through equation (3). This model will be discussed in greater detail in the next section.

III. Testlet-Based DIF

The last decade or so has seen 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" is now a general desideratum; the area of study surrounding this has been defined formally and dubbed DIF. A set of statistically rigorous and efficient procedures have been developed to detect and measure DIF. These generally fall into one of two classes; they are either based on latent variables (Thissen, Steinberg, & Wainer, 1988, 1993) or observed score (Dorans & Holland, 1993; Holland & Thayer, 1988).

Procedures for DIF studies have traditionally focused on the item. Yet, if a test is based on a broader unit [testlet] ought we not use a generalized DIF procedure that suits this broader construct? The point of this portion of the project was to utilize such a generalization and so provide estimates of the extent to which the LSAT conforms to contemporary standards of fairness.

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 in the aggregate.

Matching the Model to the Test

If a set of items were built to be administered as a unit, it is important that they be analyzed that way. There are a variety of reasons for analyzing them as a unit, but underlying them all is that if one

does not, one is likely to get the wrong answer. In an example described earlier (Wainer, Sireci, & Thissen, 1991) 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 (Gulliksen, 1950, p. 78) indicates that we would need to double the test's 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.

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 since multidimensionality is what causes DIF, we ought to 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.

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 perhaps sometimes 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 insure fairness. This can 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. As will be shown in the results section, this was the case in the operational forms of the LSAT we analyzed.

Methodology

Testlet DIF detection

The polytomous IRT model we used (specified in equation 4, above) was developed by Bock (1972). For the detection of testlet DIF we will use the well developed technology of likelihood ratio tests. The properties of these tests have been thoroughly investigated (Kendall & Stuart, 1967). Of primary importance in this work is their near optimality. Under reasonable conditions these tests are closely related to, although not necessarily identical to, the most powerful test given by the Neyman-Pearson lemma. This optimality of power is critical in situations like DIF detection, in which the testing organization would like to accept the null hypothesis; look for DIF and not find any. Not finding DIF is easily done by running poor experiments with weak statistical methods. Thus to be credible one must use the largest samples available as well as the most powerful analytic tools. In this investigation we used all of the examinees and the likelihood ratio test. Current statistical standards for DIF detection (Holland & Wainer, 1993) acknowledge that this is the best that can be done at the moment.

The basic notion of the likelihood ratio test 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*). Next fit the same model to the data allowing one testlet to have different parameters in each population (DIF) and compare the likelihood under each of the two situations. If the more general model does not yield a significant increase in the quality of the fit we conclude that the extra generality was not needed and that the testlet in question has 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 is almost exactly the one reported by Thissen et al., (1993) when we used the multiple choice model (Thissen & Steinberg, 1984) to examine differential alternative functioning (DAF). The step from DAF to testlet DIF is a small one.

Bock's 1972 Model

As described above we shall represent the trace line for score $x = 0, 1, \ldots, m_j$, for testlet j as

$$T_{jx}(\theta) = \frac{\exp[\alpha_{jx}\theta + c_{jx}]}{\sum_{k=0}^{m} \exp[\alpha_{jx}\theta + c_{jk}]}$$

The a_ks are analogous to discriminations; the c_ks analogous to intercepts. The model is not fully identified, and so we need to impose some additional constraints. It is convenient to insist that the sum of each of the sets of parameters equal zero, i.e.

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

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

$$a_{jk}=\sum_{P=1}^{P}\alpha_{jp}\left(k-\frac{m_j}{2}\right)^P \quad (5)$$

and

$$c_{jk}=\sum_{P=1}^{P}\gamma_{jp}\left(k-\frac{m_j}{2}\right)^P \quad (6)$$

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 a's are ordered the categories must be monotonically ordered as well (see the Wainer, Sireci, & Thissen, 1991, for proof). This specialization of Bock's nominal model is often referred to as Samejima's (1969) partial credit model. The polynomial representation in this application saves degrees of freedom while still providing a good representation of the data.

This version of Bock's model uses raw score within testlet as the carrier of information. While it is possible that more information might be obtained by taking into account the pattern of responses within each testlet we found that this simplification is appropriate. Moreover, basing a test scoring algorithm on number right is amply supported by general practice.

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, they 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 that 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 intra-passage correlations among the items (excess after conditioning on q). The testlet approach thus provided a more accurate estimate of the accuracy of the assessment. Through the obvious generalization, this same approach can be used to study Testlet DIF.

The basic data matrix of score patterns is shown in Table 1. In this example, there are four testlets with 10 possible scores levels each [$m_j = (10, 10, 10, 10)$]; there are a maximum of 10^4 rows. In practice there will be far fewer rows since many possible response patterns will not appear. The analysis follows what is done in item DIF situations: fitting one model allowing 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 q estimated for both groups simultaneously.

TABLE 1

Arrangement of the Data for the IRT Analyses

Testlet Score Pattern				Total Score	Frequencies	
I	II	III	IV		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}	Σ_{Sj}	f_{Ri}	f_{Fi}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
9	9	9	9	36	f_{RN}	f_{FN}

This method uses the test itself, including the studied testlet, to calculate the matching criterion. The question about whether or not to include the studied item has been carefully explored (Holland & 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_ks and c_ks 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 5 and 6). 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 that of 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.

The method of analysis utilized here departs from complex IRT analyses of the past in an important way. There is no *post hoc* 'linking' of analyses through common items and *ad hoc* assumptions about stability and linearity. Instead we specify the shape and location of the proficiency distribution. This not only provides a common metric for all comparisons (the trait in the population) but also yields estimates of the correct likelihood, thus allowing inferences about the significance of discovered effects to stand on firmer statistical ground. This methodology is based on the estimation procedure of maximum marginal likelihood (MML) developed by Bock & Aitken (1981). Thus within each analysis we estimate the item parameters as well as the distribution of the proficiency distribution of the examinees. If there is just a single group the proficiency distribution is set at N(0,1), but when there is more than a single group (as in the DIF analyses) we set one group's proficiency distribution at N(0,1) and estimate the mean of the other groups' distribution. Because this automatically adjusts for differences in the overall proficiencies of the various subgroups we can maintain all item parameters on the same scale. Moreover, we can then compare the overall performance of the various subgroups is a concise and meaningful way. We will report such a comparison among the various examinee subgroups on the LSAT as the initial part of our results section.

IV. Results

All analyses described above were performed on all four sections of two parallel forms of the LSAT. The polytomous IRT model fit well on each section. Its fit deteriorated significantly when dissimilar sections were combined. If this finding is replicated on other forms it strongly suggests that inferences made from fitting a unidimensional model to the LSAT in its entirety should be limited as much as possible, and avoided in many instances. Happily, our purposes did not require such aggregated fitting.

Our findings can be divided into three categories: 1. *overall performance* of individual subgroups on the test, 2. the *reliability* of each section, and 3. the *differential performance* of items/testlets.

1. Overall performance

Proficiency Distribution Means

	Reading Comprehension	Analytical Reasoning	Logical Reasoning (a)	Logical Reasoning (b)	Overall
Male	0.10	0.00	0.26	0.37	0.18
Female	0.00	0.00	0.00	0.00	0.00
White	0.00	0.00	0.00	0.00	0.00
African-American	-0.65	-0.92	-1.16	-1.18	-0.98
Hispanic	-0.40	-0.51	-0.69	-0.74	-0.59
Asian	-0.32	-0.19	-0.19	-0.37	-0.27
US	0.27	0.09	0.04	0.02	0.09
Canadian	0.00	0.00	0.00	0.00	0.00

Each section of the test was scaled separately within the various subanalyses. One group was assigned an arbitrary mean level of performance of 0.0 and the other group's performance was characterized relative to that. All figures are in standard deviation units; the distribution of scores within groups can be thought of as Gaussian. Thus we see that men scored about .1σ above women in Reading Comprehension, were essentially identical to women in Analytical Reasoning, and were about a third of a standard deviation higher in Logical Reasoning. On average (weighting all sections equally) men scored about two-tenths of a standard deviation higher than women.

Obviously, any recommendations made on the basis of the findings from the analyses of only these two LSAT forms (which are really the same form with the presentation order of the sections shuffled), must be confirmed on other forms before being considered for implementation. These results are suggestive (albeit sometimes strongly so), not conclusive.

2. *Reliability*

Group Reliabilities

	Reading Comprehension	Analytical Reasoning	Logical Reasoning (a)	Logical Reasoning (b)
Male	0.69	0.60	0.79	0.77
Female	0.69	0.60	0.77	0.77
White	0.69	0.58	0.78	0.75
African-American	-0.68	-0.56	0.69	0.79
Hispanic	-0.69	-0.60	0.75	0.72
Asian	-0.71	-0.59	0.77	0.77
US	0.68	0.59	0.79	0.77
Canadian	0.68	0.59	0.79	0.77
Overall	0.71	0.60	0.79	0.79

Shown above are the reliabilities of each section of the LSAT within the various subgroups specified. As can be seen there is very little variation in section reliability across subgroup and so one can usefully use the single overall value shown at the bottom. What differences there are in the section reliability among groups is largest in the Logical Reasoning section.

Shown below is a comparison of these reliabilities with those obtained using traditional methodology (labeled Item-Based Reliability' in the table) that does not take into account the testlet structure of the LSAT.

Reliability of the LSAT

Section	Item Based Reliability	Testlet Based Reliability	Reliability Range	Length Increase Required
Reading Comprehension	0.79	0.71	0.68-0.71	1.54
Analytical Reasoning	0.76	0.60	0.56-0.60	2.11
Logical Reasoning (a)	0.77	0.79	0.69-0.79	0.89
Logical Reasoning (b)	0.77	0.79	0.72-0.79	0.89

The column labeled "Reliability Range" reflects the variability of reliability seen across all subgroups (reported explicitly in the table that immediately preceded this one). The column labeled "Length Increase Required" is an estimate, obtained from the Spearman-Brown Prophesy formula, of how much longer that section would have to be made in order for it to have a testlet-based estimate of reliability that is equal to the estimate obtained through traditional test theory when the latter does not explicitly model the local dependence observed. Analytical Reasoning would have to be more than doubled, yielding eight situations instead of the current four. Similarly Reading Comprehension would require six passages rather than the current four to yield the reliability of .79 that was previously claimed.

These findings may have implications for the way that the various test sections are combined to yield total score. If all sections are equally reliable there are psychometric arguments in support of counting each section equally. The nominal reliability of the four sections examined are sufficiently close to warrant this approach. However, if we accept the testlet-based estimates of reliability all sections are no longer equally reliable. If we want the composite score to be maximally reliable we must either weight the various sections differentially (Wang & Stanley, 1970) or make the various component sections equally reliable. The latter can be accomplished in at least two ways; adjusting the length of the sections, or using ancillary information in the estimation of each section's score.

There are two observations that can be made from these results. First that explicitly modeling the testlet structure of the test provides us with a more accurate estimate of the parallel forms reliability of each section. This results in a decrease in those two sections of the test that are substantially clustered. Second, the reliability of the Logical Reasoning sections, which does not have much clustering (there are two pairs of items), has actually increased a little bit. This increase is due to the use of IRT which extracts more information from the test response pattern of each individual examinee than does using merely the total number correct.

3. *Differential Performance*

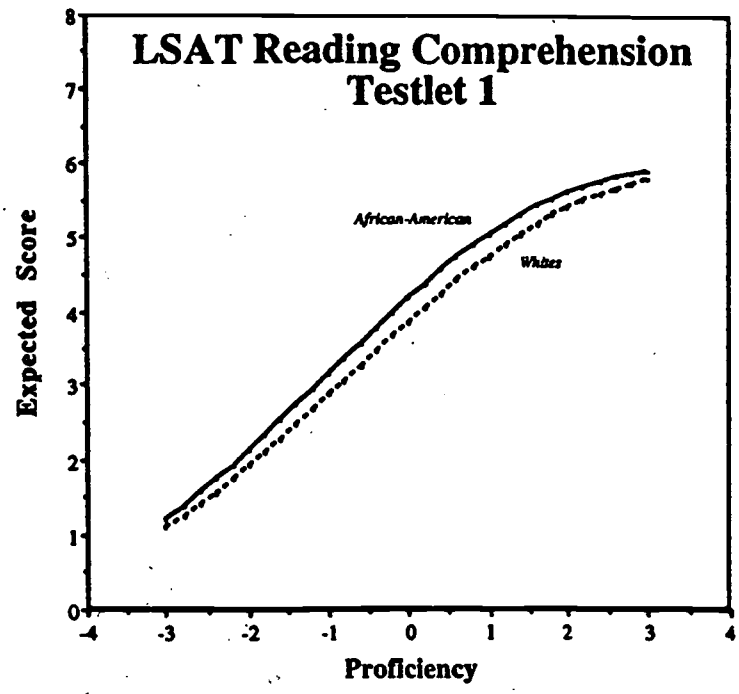
All of the individual items were screened for differential performance (DIF) using Mantel-Haenszel methods during *pro forma* item analysis. Very few were found to deviate from acceptable standards of fairness. However, when these individuals items were viewed from the point of view of a testlet a different pattern emerged. As an example of what we found consider the Mantel-Haenszel statistics associated with the comparisons between Whites and African-Americans (below). The standard interpretation of the Mantel-Haenszel (MH) statistic (see Dorans & Holland, 1993, p. 42) classifies items as problematic if $|MH|>1.5$. Under this rule only one item (#3) is questionable. The sign of the statistic indicates which group is being disadvantaged; a positive value means it favors the focal group (African-Americans in this instance), a negative value the reference group (Whites). It would not be surprising to find that these two passages and their associated items passed muster after

pre-test screening (in those cases where there was enough data to allow it), since only one item stuck out over the statistical barrier against discrimination, and that one only barely. However if we look at these statistics as a group, clustered by passage there is an obvious pattern. Virtually all of the items associated with passage 1 favor the focal group (total MH = 4.82) and 5 of the 7 items associated with the fourth passage favor the reference group (total MH = -1.38).

Passage	Item Number	Mantel-Haenszel Statistic
1	1	0.88
	2	-0.04
	3	1.59
	4	0.98
	5	0.53
	6	0.88
	Total	4.82
4	22	0.26
	23	-0.57
	24	-0.42
	25	-0.31
	26	-0.54
	27	0.56
	28	-0.36
	Total	-1.38

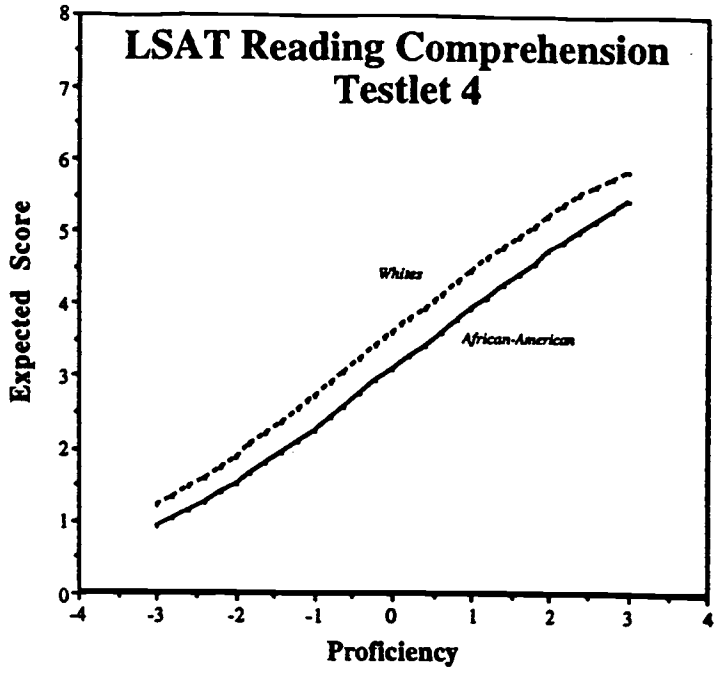
The results obtained from the traditional DIF analysis were suggestive. When the data were analyzed for DIF with the testlet structure explicitly modeled we found that at all levels of proficiency African-Americans had a higher expected score on this item than Whites. The extent to which this is true is shown in Figure 1, below.

Figure 1



On the other hand, Whites are advantaged on passage 4 (Figure 2).

Figure 2



Thus using a test scoring procedure that explicitly models the structure of the LSAT we were able to detect and measure possible deviations from fairness that were hidden previously. Our principal substantive finding, after examining DIF for all testlets on five focal groups (women, African-Americans, Hispanics, Asian-Americans, and Canadians), is that there was very little DIF associated with any particular testlet that was consistent across most of the focal groups, other than the two shown above. Both Passage 1 and 4 of the Reading Comprehension section were found to consistently show differential performance.

We have used a simplifying display method, which, though not particularly arcane, may be unfamiliar. Thus we shall make a brief aside to comment on it. Dichotomous IRT traditionally represents an item's performance with the traceline associated with the probability of a correct response; a plot of p(correct) vs. θ. There are really two tracelines for a dichotomous item. The one not plotted is associated with the probability of getting the item wrong (1-p). In the dichotomous case the two tracelines provide redundant information and therefore, in the interests of parsimony, are rarely both presented. In the polytomous case there are more than two tracelines. There is one traceline associated with being in each category. This is often presented as a sequence of tracelines. When the number of score categories is greater than 3 or 4 this practice often results in a web of lines that defy easy interpretation. To ease this problem we plot the expected score curve. These curves are usually well-behaved and quite regular in appearance. They are also in a metric that is meaningful to the observer. Expected score is calculated by choosing a large number of points on the proficiency axis (say 20), and at each point calculating the expected score. This is done by multiplying the value of each score category by the probability of obtaining that score at that value of θ. The score category values are determined by the test developer; the probabilities are estimated from the model (equation 4). In DIF analyses we use different values of the parameters for the focal and the reference groups to estimate these probabilities. We have found that, because of trade-offs in the model it is almost impossible to judge the severity of the DIF through

examination of the absolute differences in the parameter values. But when these differences are characterized in the metric of expected score our understanding is improved substantially.

A testlet examination of White-African-American performance on Reading Comprehension: A brief case study

We have reported the most substantial findings that have been uncovered. In this subsection we shall step through one set of analyses, for one test section for one focal group, to indicate how we arrived at these final results. It would be wasteful of effort to repeat this explanation for all sections and all focal groups. The details of all of those analyses are contained in the appendices in the attached notebooks and, in electronic form on the enclosed disks. Also on those disks are all of the associated files required to duplicate any of the sets of outputs.

Step 1: Fit the joint data for White and African-American examinees with a single model

The Reading Comprehension section of these forms of the LSAT consists of four testlets. Each testlet is comprised of a passage and (6, 7, 8, and 7) items respectively. Because of the possibility of a zero score this means that the scores achievable on each testlet fell into (7, 8, 9, and 8) score categories respectively. Thus (referring back to Table 1) each group had the possibility of 4,032 (=7x8x9x8) score patterns. The initial analysis we did treated the two groups of examinees as one and estimated the mean of the proficiency distributions for those two groups as well as estimating the fit.[4] Because we treated both groups as one, we were explicitly assuming no DIF. We found that -2loglikelihood for this model was 8,224. After fitting this model there were 33 free parameters. We found that a preliminary estimate of the mean of the White proficiency distribution was .8σ higher than that for the African-Americans. Part of this difference may be due to testlet DIF. We will amend this figure as the analysis progresses.

Step 2: Relax the equality constraints completely and fit a more general model

In this model we allowed the parameters for each of the testlets to be estimated separately for each of the two groups; ALL DIF run. This was done in a single run in the same way as Step 1, by releasing the equality constraints on the parameters while simultaneously fixing the proficiency distributions to be the same as those estimated in Step 1. This yielded a -2loglikelihood of 7,408 but with 64 free parameters.

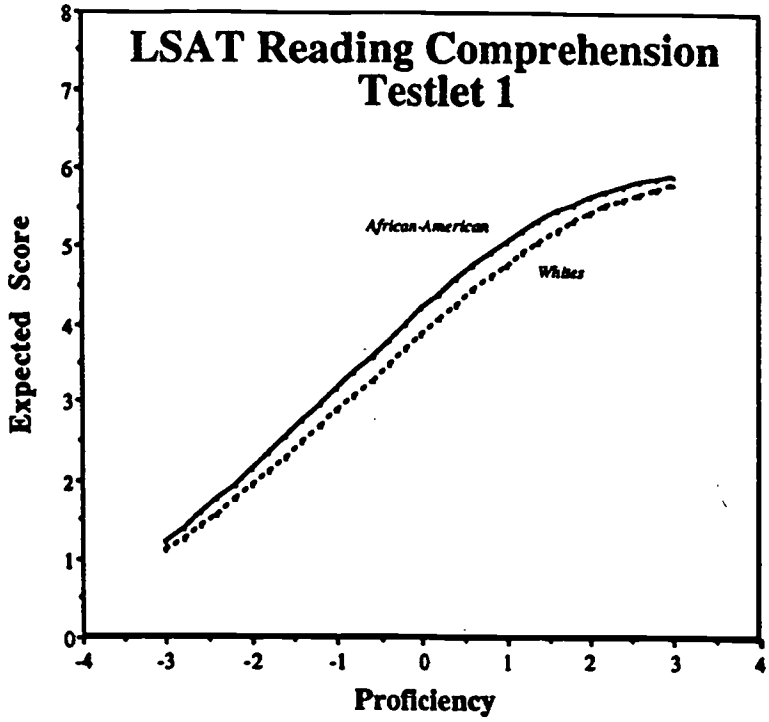
Step 3: Compare NO DIF with ALL DIF

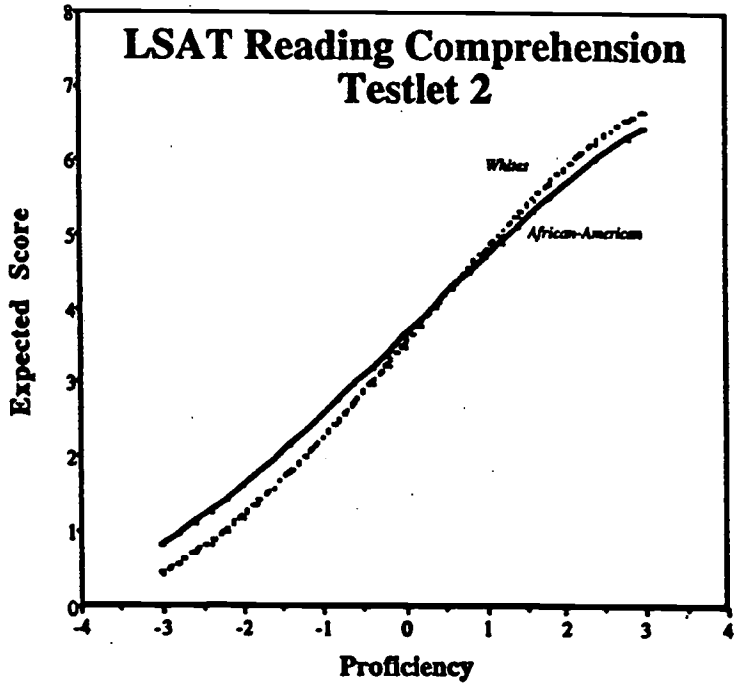
Subtracting the fits of the two runs obtained in steps 1 and 2 resulted in a χ^2 of 816 (8,224—7,408) on 31 (64—33) degrees of freedom. This is statistically significant by any measure. Thus we can reject the NO DIF model. The task was then to track down where, which testlet(s), was the DIF.

[4]As a brief operational aside, we ran this model as if there were 8 testlets, but Whites omitted the last four and African Americans omitted the first four. Then we applied equality constraints to set testlet 1 = testlet 5, testlet 2= testlet 6, testlet 3 = testlet 7, and testlet 4 = testlet 8. This effectively fits a single model of NO DIF while allowing separate estimates of proficiency to be obtained for each group. It also provides a graceful way to fit successively more general models.

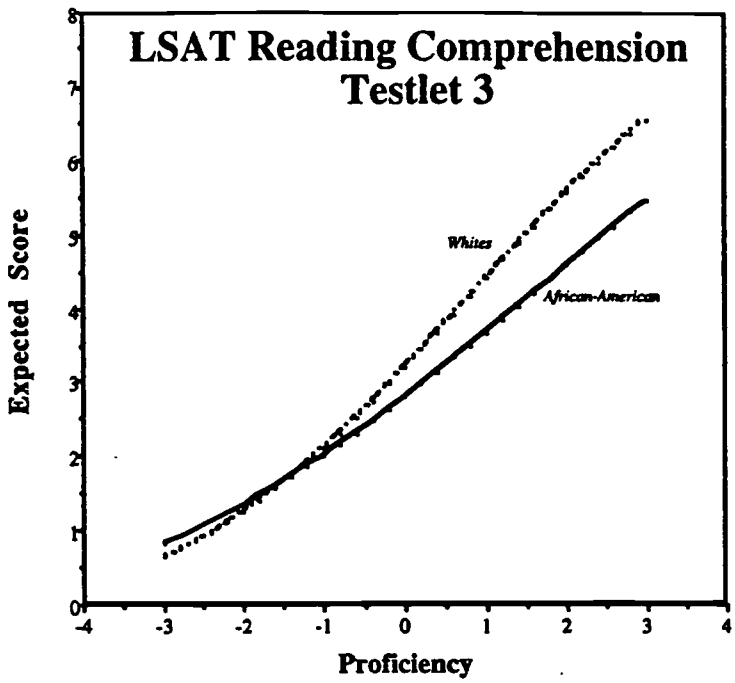
Step 4: Exploratory analysis - plot the tracelines from the ALL DIF analysis.

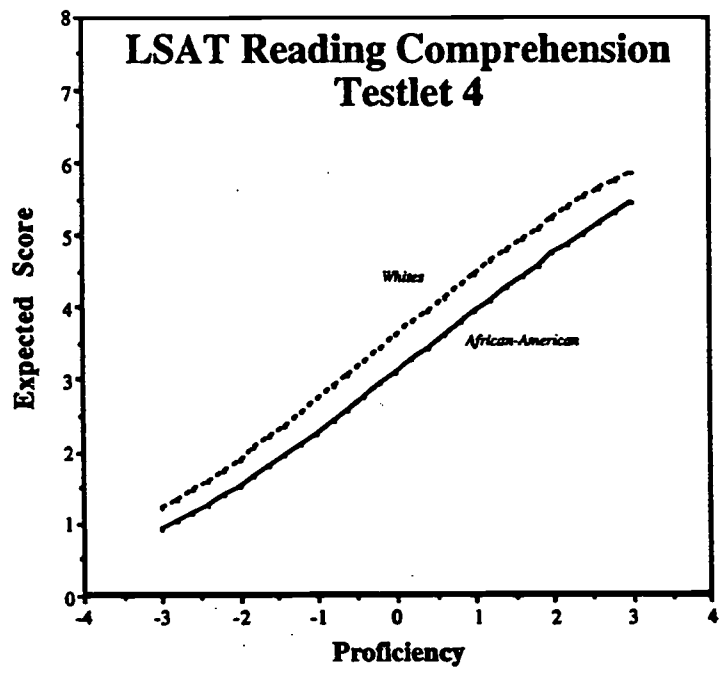
From the finding in Step 3 we knew that at least one and at most four testlets had White/African-American DIF. Tracking down which one(s) is easier with some exploration. We plotted the expected score curves for each testlet. From these we observe that there is very little difference between the expected score curves for testlets 1 and 2.





As compared to more substantial differences between the groups for testlets 3 and 4.





This observation led us to the next step.

Step 5: Fit a model that sets testlets 1 and 2 equal in the two groups, but allows testlets 3 and 4 to be unequal.

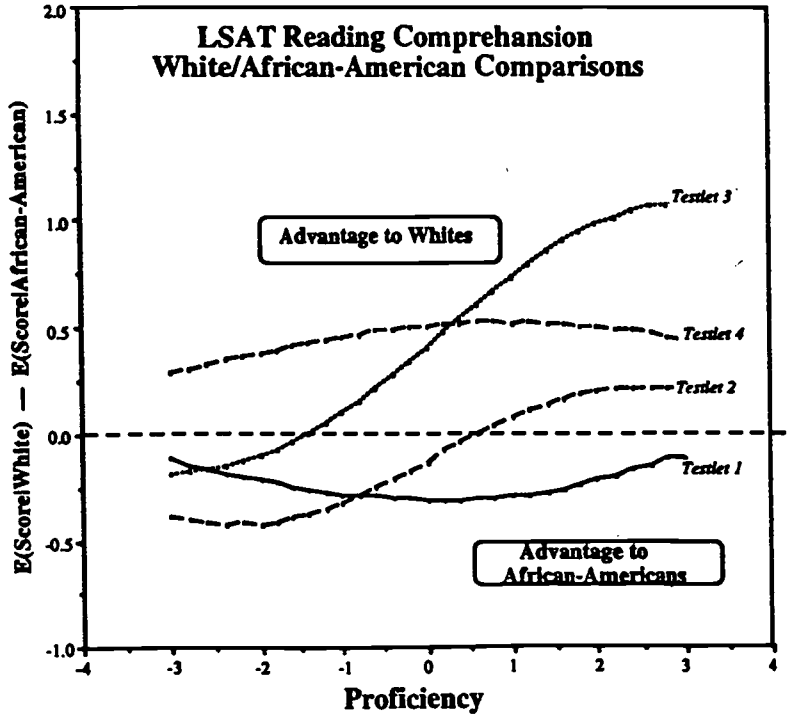
When this model was fit (3 & 4 DIF) we found that -2loglikelihood was 7,528 and there were 50 free parameters. This model fits much better than the NO DIF model (χ^2 of 696 on 17 degrees of freedom) but is still significantly worse than the ALL DIF case (χ^2 of 120 on 14 degrees of freedom). Inspection of the previous figures led us to fit a model in which only testlet 1 was fixed to be equal in the two groups.

Step 6: Fit a model with testlets 2, 3 & 4 unequal.

When this model was fit the -2loglikelihood was 7,422 with 58 free parameters. This represented a significant improvement over the previous model (step 5) with a χ^2 of 106 on 8 degrees of freedom) and it was not seriously worse than the ALL DIF model (χ^2 of 14 on 6 degrees of freedom). It was this model that we chose to represent the data. This model was fit without fixing the means of the proficiency distributions and the estimates of those means thus obtained indicted that Whites were .65σ higher than African-Americans. Note that this is a smaller difference than was observed in the original NO DIF run. The reason for this is that some of the difference between the two groups was absorbed by allowing the three of the testlets to have different parameters. So by adjusting for the DIF a smaller estimate of the between group difference obtained.

There are two details that deserve mention. First, although the likelihood ratio comparing ALL DIF with 2, 3, & 4 DIF models is modest, compared to the other ones, it still achieves a nominal value of significance. This ought not be followed slavishly, for with large enough sample sizes any observed difference will be 'statistically significant'. In this study we had a Focal group of 3,149 and a Reference group of 33,681. With such sample sizes very small differences are detectable. The absolute size of the difference between the two curves is of greater importance. Note that the tiny difference observed between the expected score curves on testlet 1 could be detected.

A summary of the differences observed is shown below:



From this summary we can see that since most of the Focal group is concentrated between -1.5 and .5 on the proficiency dimension, the DIF in testlet 4 has the most profound negative effect on African-Americans. Note that this negative effect is more than compensated for by the summed positive effects of the other testlets. This happy outcome is obscured by the apparent large negative DIF observable at high levels of proficiency. But because of the location of African-American proficiency distribution this negative effect is substantially ameliorated.

In the Table below is a summary of the fit statistics that characterized these analyses. This Table also includes the (1, 3 & 4) DIF analysis that was not helpful.

LSAT Reading Comprehension
Summary of White/African-American DIF

Model	-2loglikelihood	Number of Free Parameters
No DIF	8,224	33
3 & 4 DIF	7,528	50
1, 3, & 4 DIF	7,438	57
2, 3, & 4 DIF	7,422	58
All DIF	7,408	64

This concludes the DIF analysis for one Focal Group. The same procedure was followed for four other focal groups (Hispanics, Asian-Americans, Women, Canadians). In each case we did exactly the same thing to arrive at a best-fitting model. Of course, for each analysis we arrived at different results. For African-Americans testlets 3 and 4 may have been the passages with the most DIF. For other focal groups other passages stood out. Is there some way to decide the worth of each passage over all of the focal groups? Given the spirit of DIF, looking at performance within a particular group, is such a question sensible?

On the off-chance that such a formulation is required we summarized the total DIF for each focal group by summing the area between the expected score curves, weighted by the proficiency distribution of the focal group (Wainer, 1993). Because this methodology is both obvious and unique it is worthwhile taking a short diversion to explain it.

Suppose, in addition to detecting and representing the DIF in a testlet we also wish to measure its amount. The following is a development of a model-based standardized index of impact that grows naturally out of the IRT approach taken in this project. The basic notion supposes that the impact of item *i* is the difference between the expected score on that testlet if one is a member of the focal group versus what would have been the score if one was a member of the reference group (i.e., suppose one went to sleep *focal* and awoke *reference*; how much does that change one's expected score?).

To answer this we need a little notation:
Define:

$E_F(x_i \mid \theta)$ be the expected score on testlet i for a member of the focal group, conditioned on having proficiency θ, and

$E_R(x_i \mid \theta)$ be the expected score on testlet i for a member of the reference group with the same proficiency.

Further, let us define proficiency distributions for each group as $G_F(\theta)$ and $G_R(\theta)$ respectively.

Last, suppose there are N_F and N_R individuals in each group. Now we can get on with the derivation of the indices of standardized impact. When there is DIF and a person changes group membership overnight their expected score on the item changes. Thus the amount of impact is

$$E_F(x_i \mid \theta) - E_R(x_i \mid \theta)$$

But, this must be weighted by the distribution of all of those affected, specifically, by $G_F(\theta)$. Thus, we will define the standardized index of impact, T(1) as

$$T(1) = \int_{-\infty}^{\infty} [E_F(\chi_i|\theta) - E_R(\chi_i|\theta)] dG_F(\theta)$$

Obviously, this is the average impact for each person in the focal group. Note that this is bounded regardless of the model used to generate the expected values This index comes close to characterizing what we want. Indeed, for many purposes this may be just right.

Is this always the right thing to be looking at? The amount of a testlet's impact depends on the choice of the focal group. An item might be unjust to one focal group, but just fine for another. Thus, it might be important for purposes of comparison to have a measure of *Total Impact,* or $T(2) = N_F T(1)$, or

$$T(2) = N_F \int_{-\infty}^{\infty} [E_F(\chi_i|\theta) - E_R(\chi_i|\theta)] dG_F(\theta)$$

The concept of *Total Impact* is often a useful one in test construction. Consider that one constraint in test construction might be to choose a subset of testlets from a pool such that the total impact is minimized. Dorans (personal communication) describes a "melting pot" reference population, which is made up of all of the various focal groups. He suggested that one might then calculate the Total Impact (using an index like T(2)) for each group relative to the whole. The operational testlet pool might be the one of requisite size that minimizes Total Impact (summed over the entire examinee population).

The concept of Total Impact allows us to consider the situation in which one testlet has only a small amount sex *DIF* but no Aztec *DIF,* whereas another testlet might have more Aztec *DIF* but no sex *DIF.* This method allows us to choose between them based upon their total effect (sex groups are typically much larger than the number of LSAT-taking Aztecs)[5].

Shown in the table below is the standardized impact for each testlet for each focal group. Testlet 1 is the one that seems to disadvantage these groups most uniformly, whereas Testlet 4 is the most disadvantageous for the reference groups. The figures provided under each focal group are what we have

[5]Another variation on this theme was suggested by Nambury Raju (personal communication, August 16, 1989). He suggests that it might be more useful to have a measure of *Proportional Impact* rather than *Total Impact.* To calculate this we need only replace "N_F with N_F/N, where N refers to the size of the melting pot reference population." I am fond of this idea, but it remains for experience with such indices in practical situations to inform us more fully of their worth.

defined as T(1). These are summed across focal groups in the column labeled "unweighted means." The last column "Focal weighted means" is the sum of the T(2) statistics; the weights used are at the bottom of the table. From this column we see a more precise measure of the effect of testlet DIF. It averages about one-fourth of a point against the various focal groups on Testlet 1 and about four-tenths of a point in favor of the various focal groups on Testlet 4.

Mean Standardized Impact

	Focal Group						
Testlet	Females	African-American	Hispanics	Asians	Canadians	Unweighted Means	Focal Weighted Means
1	-0.3	-0.3	0.0	-0.1	-0.1	-0.2	-0.24
2	-0.1	-0.1	0.0	-0.2	-0.2	-0.1	-0.11
3	0.1	0.4	0.3	-0.2	-0.3	0.2	0.09
4	0.4	0.5	0.3	0.2	0.3	0.3	0.38

	Weights (N's)				
Focal	18,716	2,971	2,103	2,633	2,294
Reference	22,022	32,778	32,778	32,778	41,302
Total	40,738	35,749	34,881	35,411	43,596

This completes the analysis for the Reading Comprehension section of these forms of the LSAT when comparing White and African-American examinees. For the purposes of all of these analyses the two forms examined were merged by ignoring order effects and joining data from identical testlets. In the attached notebooks, which should be thought of as appendices to this report, are parallel results for the other four focal groups, as well as for the other three sections of the test. Printed versions are enclosed for Analytical Reasoning, which does have a testlet structure. Only the electronic format is included for the two Logical Reasoning sections, since there is almost no testlet structure in those sections, and hence those results are only superficially different from what would be obtained from traditional, item-based, analyses.

V. Conclusions

This examination has shown that the testlet structure of the Reading Comprehension and Analytical Reasoning sections of the LSAT has a significant effect on the statistical characteristics of the test. The testlet-based reliability of these two sections obtained in these analyses is considerably lower than what was previously calculated under the inaccurate assumption of local conditional independence. We believe that this discovery should change current practice. At a minimum, any reporting of section reliability ought to be modified to reflect our current knowledge. Moreover analyses of current and future forms of these sections of the LSAT ought to model explicitly the testlet structure before calculating section reliability. Because of the test's overall length we don't believe that the total test reliability that is reported is far enough wrong to have serious need for correction.

A second possibility would be to boost the section reliabilities up to the levels that were previously thought to hold through one of two modifications: (a) increase the length of the two testlet sections (6 passages in Reading Comprehension and 8 scenarios in Analytical Reasoning), or (b) use such statistical procedures as empirical Bayes estimation to 'borrow' reliability for each section from the other sections. Choice (a) would cause a substantial increase in testing time; choice (b) is practically free. Our experience with the use of procedures of this sort leads us to believe that the use of other sections to bolster the section scores would be substantial enough to yield section scores of sufficient statistical stability to justify reporting them separately. This may be a welcome addition to the LSAT for both Law Schools and examinees. Since the marginal cost of doing this is essentially nothing, we believe that it is an idea worth serious and immediate consideration.

The DIF analyses provide reassurance. The size of the differences in the expected score curves that was detected as significant was very small. This assures us of the impressive statistical power of this methodology. The size of these differences in the worst case was small enough to suggest that no serious problems of fairness exist. Moreover, even these small differences were reduced to almost nothing through the balancing across testlets. This balance cannot be swallowed whole. Because performance on the test section itself determined the stratifying variable, the overall balance (zero overall DIF) is almost tautological. That the balancing works as well as it does at all levels of examinee proficiency is not mathematically determined. It is one sign of a fair test.

Is this all? No. It is disquieting to note that the Reading Passage with the greatest impact on all focal groups was one whose general topic was constitutional law. If ever there was a population of examinees for whom this topic is a fair one to include on a test it is the LSAT population. Yet it shows up as the worst one. Why? One explanation is that it is a perfectly fair item. Perhaps it is the other three passages that have DIF in favor of the various focal groups. Since we are using an internal criterion to measure DIF (to stratify examinee performance) such a conclusion would yield the observed results. The only way to determine which of these two hypothetical interpretations is more nearly correct is through the use of more information. The most obvious source of information would be validity studies. One prospective study would be to see how predictive of success in law school are each of the passages. One can analyze such data in a way faithful to the structure of the test by using law school grades as the stratifying variable in a DIF study. Then see which of the passages shows DIF and in which direction. When a validity criterion is used as the stratifying variable in a DIF study, we are no longer studying DIF; we are studying bias (in both the statistical and the pejorative sense). We believe that such a study, given that validity data are in the process of being gathered and coded, would be easy to do and would yield important and useful information.

> *"What we observe is not nature itself but nature exposed to our method of questioning."*
> Werner Heisenberg (1958).

This investigation, based as it was on the testlet structure of half of the LSAT and polytomous IRT models, provides a glimpse into what is achievable with modern psychometric tools. In Heisenberg's sense, we are using a more delicate and more rigorous method to question our data. Using such tools has given us a deeper understanding of the structure of the test and of its measurement characteristics. In addition to the 'bad news' that two of the test's sections appear to be considerably less reliable than was previously thought, we found that through a more efficacious weighting of items, the Logical Reasoning sections were more reliable that we thought. The use of these sorts of modern measurement models can allow us to suck out as much information as there is within each examinee's responses. Of at least equal importance, they also provide a more accurate estimate of the error that remains. These procedures have not yet become widely used operationally, and so despite their theoretical advantages, we do not have the same enormous experience with them as has accumulated with traditional procedures. However, with added experience testlet-based IRT procedures show great promise.

VI. References

American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association.

American Psychological Association (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Anastasi, A. (1961). *Psychological testing* (2nd edition). New York: Macmillan.

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. (1993). 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. Pp. 3-23.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores* (Pps. 392-479). Reading, MA: Addison-Wesley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

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. (1993). 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. Pp. 115-122.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179-197.

Brown, W., & Thomson, G. H. (1925). *The essentials of mental measurement* (3rd edition). London: Cambridge University Press.

Crumm, W. L. (1923). Note on the reliability of a test, with a special reference to the examinations set by the College Entrance Board. *The American Mathematical Monthly, 30(6),* Sept.- Oct. 1923.

Donlon, T. F., & Angoff, W. H. (1971). *The Scholastic Aptitude Test.* In W. H. Angoff (Ed.) *The College Board Admissions Testing Program* (pp. 15-47). New York: College Entrance Examination Board.

Dorans, N. J. & Holland, P. W. (1993). "DIF detection and description: Mantel-Haenszel and standardization." Chapter 3 in P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Pp. 35-66.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

Guilford, J. P. (1936). *Psychometric methods* (1st edition). New York: McGraw-Hill.

Gulliksen, H. O. (1950/1987). *Theory of mental tests,* New York: Wiley. (Reprinted in 1987 by Lawrence Erlbaum Associates; Hillsdale, NJ).

Haberman, S. J. (1978). *Analysis of qualitative data. Volume I: Introductory topics*. New York: Academic Press.

Heisenberg, W. (1958). *Physics and Philosophy*. New York: Harper & Row.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577-601.

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

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holzinger, K. (1923). Note on the use of Spearman's prophecy formula for reliability. *The Journal of Educational Psychology, 14,* 302-305.

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.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crumm's criticism. *The Journal of Educational Psychology, 15,* 193-204.

Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics*, Volume II. London: Charles Griffin.

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, Mass.: Addison Wesley.

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

Neyman, J., & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 174-240 and 263-294.

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,* Part 2, No. 17.

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

Thissen, D. (1990). *MULTILOG (version 6.0) user's guide*. Mooresville, IN: Scientific Software.

Thissen, D. (1991). *Multilog user's guide* [computer program]. Chicago, IL: Scientific Software.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy & I. Bejar (Eds.) *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 79-97.

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

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

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* (Pps. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). 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. Pp. 67-113.

Thorndike, R. L. (1951). In E. F. Lindquist (Ed.) *Educational measurement. Washington, D.C.*: American Council on Education.

Tierney, L. (1990). *LISP-STAT: An object-oriented environment for statistical computing and dynamic graphics*. N.Y.: Wiley.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*, 37-45.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Wainer, H. (1993). 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. Pp. 121-135.

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

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer, H., Sireci, S. G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Wang, M. D., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. Princeton, NJ: Educational Testing Service Research Report RR-89-41.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Yen, W. M. (1992, April). *Scaling performance assessments: Strategies for managing local item dependence*. Invited address presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca.

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

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").